Universidad de la República
Facultad de Ciencias Sociales
DEPARTAMENTO DE ECONOMIA

Documentos de trabajo

Symbolic time series analysis in economics

J. G. Brida

Documento No. 10/00
Diciembre, 2000

Symbolic Time Series Analysis in Economics

Juan Gabriel Brida

Symbolic Time Series Analysis in Economics

Abstract

In this paper I describe and apply the methods of Symbolic Time Series Analysis (STSA) to an experimental framework. The idea behind Symbolic Time Series Analysis is simple: the values of a given time series data are transformed into a finite set of symbols obtaining a finite string. Then, we can process the symbolic sequence using tools from information theory and symbolic dynamics. I discuss data symbolization as a tool for identifying temporal patterns in experimental data and use symbol sequence statistics in a model strategy. To explain these applications, I describe methods to select the symbolization of the data (Section 2), I introduce the symbolic sequence histograms and some tools to characterize and compare these histograms (Section 3). I show that the methods of symbolic time series analysis can be a good tool to describe and recognize time patterns in complex dynamical processes and to extract dynamical information about this kind of system. In particular, the method gives us a language in which to express and analyze these time patterns. In section 4 I report some applications of STSA to study the evolution of different economies. In these applications data symbolization is based on economic criteria using the notion of economic regime introduced earlier in this thesis. I use STSA methods to describe the dynamical behavior of these economies and to do comparative analysis of their regime dynamics. In section 5 I use STSA to reconstruct a model of a dynamical system from measured time series data. In particular, I will show how the observed symbolic sequence statistics can be used as a target for measuring the goodness of fit of proposed models.

1 Introduction

The study of economic systems is different from those in laboratories. In contrast to laboratory experiments, it is generally not possible to influence the medium under consideration or to repeat such observations under the same conditions. This often leads to rather short and noisy observational data, which are sometimes irregularly sampled or show a non-stationary behavior. Available time series give often just one single observation per year. Because of the difficulties exposed above, well-known methods of linear data analysis (like correlation and spectral analysis) and of nonlinear dynamics (like estimates of fractal dimensions or Liapunov exponents) are inadequate to characterize patterns when we work with high-dimensional systems. This motivates us to try to find other analytical methods that work better in our context.

Economist have long been concerned with the explanation of business cycles, the literature on the subject is wide extensive and the number of theories that is vast. However, many of the models that have been proposed to explain economic fluctuations can be grouped into two main categories: linear Gaussian random processes and non-linear dynamics. Both of these approaches assume that business cycle models should include both deterministic and stochastic components. In the first approach, that we can associate with the seminal works of Slutsky and Frisch and can be named the "econometric approach", it is assumed that the deterministic element is inherently linear and that the fluctuations are produced by random shocks. Nonlinear models differ from the above because they explain fluctuations by the presence of non-linearities in the fundamental functions of the model and maintain that random shocks play no essential role in the behavior of the dynamical system. The two approaches described above are distinctly different, but it is not clear which is the better one. It is also clear that one can also think on a large variety of intermediate situations, where there is not a great predominance of the deterministic or the stochastic part of the models. Pure deterministic (non-linear) and pure stochastic models are the extremes that one can take as reference but from real data analysis, it seems that neither of these possibilities is the most appropriate.

In which follows I will attempt to use a new approach that arises naturally in the context of data analysis and processing. This method is known as Symbolic Time Series Analysis (STSA) or data symbolization. In order to recognize time patterns in complex dynamical processes we need a language to express and analyze these patterns and data symbolization seems to be a very effective method to introduce such a language. The idea behind data symbolization is simple: the values of a given time series data are transformed into a finite set of symbols obtaining a finite string. That is, we have to substitute actual signals with their symbolic representation. The approach, introduced in Lehrman, Rechester and White (1997)**,** Tang and Tracy (1997), Tang, Tracy and Brown (1997) and Tang, Tracy, Boozer and Brown (1995), has several successful applications to complex time series data in different areas like physics, astronomy and engineering[1] and could be a contribute to the debate in economic fluctuations. We build on this previous work and adapt known methods to economic time series data.

[1] See Armfield, Daw, Durbetaki, Finney and Green (1998), Armfield, Daw, Dralimeier, Durbetaki, Finney, Green, Kennel and Wagner (1998), Connolly, Daw, Finney and Kennel (1998), Daw, Finney and Green (1998), Daw, Finney, Nguyen and Halow (1998a), Daw, Finney, Nguyen and Halow (1998b) and Daw, Edwards, Finney and Nguyen (1998) for applications in the study of cycle variability, control of period doubling bifurcation values in internal combustion engines and other measurements of engineering systems, Cruchfield and Young (1989) for applications to characterize complex signals, Lehrman, Rechester and White (1997) for applications on the construction of finite state models and detection of dynamical correlation, Tang, Tracy, Boozer, deBrown and Brown (1994) for applications in reconstruction of chaotic signals, Witt, Kurths, Benz and Schwarz (1993) for applications to astronomy data.

Resent research in dynamical systems theory has revealed that time irreversibility[2] is a critical feature that can be used to ascertain which of the above approaches works better. Time irreversibility means that the temporal statistical properties of a measurement series differ when considering the forward and backward time realizations. Time reversibility is inherent to random Gaussian linear processes while non-linear dynamical models are inherently irreversible; see Diks, van Houwelingen, Takens and DeGoede (1995) and Weiss (1975) for detailed expositions about this field. So, by evaluating time irreversibility of experimental data, we may have an idea about the kind of functional dependence of the data. For the experimental data analyzed in Armfield, Daw, Dralimeier, Durbetaki, Finney, Green, Kennel and Wagner (1998), data symbolization appears to be a very effective method for detecting and quantifying irreversibility.[3] The data, the history and the approaches that appear in modeling cyclic combustion variability in spark-ignited engines are akin to those that we have in business cycles and so it is likely that symbolic time series analysis help to quantify time irreversibility in economic time series data.

Symbolic time series analysis is also concerned with well-defined measures of uncertainty and complexity like Shannon entropy and algorithmic complexity. Given a data set, these measures gives us some guidelines to determine what kind of models are admissible for the data in terms of determinism and complexity requirements. Maximal values of these measures characterize pure stochastic processes while minimal values are often related to simple (no chaotic) deterministic systems. But, in general, actually data present intermediate levels of uncertainty and complexity. Then, one powerful application of STSA is to contrast simulated data obtained from a proposed model with real data and quantify if differences are statistically significant or not.

Briefly, STSA consists in the transformation of a given time series into a symbolic sequence characterizing the former only through a few distinct symbols. The quantization is made in the following way: if we have a sequence of data $\{x_1, x_2,\ldots, x_t,\ldots, x_T\}$, where $x_t \in R^q$ for $t = 1, 2,\ldots, T$ (i.e., we have T vectors of data each one q-dimensional), we divide R^q in n pieces and label the pieces with symbols (usually the numbers $1, 2,\ldots, n$ or $0, 1$ if we make a binary partition). Then we transform the sequence of data $\{x_1, x_2,\ldots, x_t,\ldots,x_T\}$ into the sequence of symbols $s_1s_2\ldots s_t\ldots s_T$, where $s_t = s$ if and only if x_t belongs to the piece labeled with s. For example, if we have a binary partition in a one-dimensional space, we assign to each data value the symbol *"0"* or *"1"* according to whether the data value is below or above a given threshold, respectively. The division in pieces will depend on the type of data we have, to agree with an underlying model or some economic criteria. Once we have the resulting symbolic sequence, we analyze it looking for regularities to describe the dynamical behavior of the system.

In order to recognize time patterns in complex dynamical processes we need to have a language in which to express these patterns and we can do it using data symbolization. The main property of data symbolization is the fact that no assumptions about the structure of the underlying dynamical system must be stated; that is, the method applies for deterministic or stochastic, linear or non-linear systems.

STSA derives from symbolic dynamics, an important branch of dynamical systems. The methods of symbolic dynamics have been applied in dynamical systems theory to describe the behavior of non-linear systems when a Markov (or generating) partition is known. Via conjugation,

---

[2] Time irreversibility is defined such that a qualitative or quantitative description of a time series and its time-reversed realization differ significantly. It is known as an indicator of non-linear structure and data with the time irreversibility property may not be explained as a Gaussian linear random process.

[3] See Voss and Kurths (1998) for an alternative method to discriminate between symbol sequences that can be described by linear stochastic processes and non-linear dynamical systems.

a particular dynamical system is transformed into a shift of finite type. Then, to analyze the shift, all the apparatus of symbolic dynamics theory can be used. Famous examples of this kind of applications are the horseshoe of Smale and the logistic map (particularly used in economics almost like the "unique" example of chaotic behavior). In present days symbolic dynamics is used in a variety of fields including ergodic theory, topological dynamics, complex dynamics, information theory and coding. Although the method of data symbolization is inspired by symbolic dynamics theory, it is not equivalent because for doing the partitions we don't know the system that generates the dynamics and we can also have the presence of noise.[4] Then, the methodology is inspired in symbolic dynamics but in place of theory it must rely on empirically and heuristically motivated arguments.

2 Symbolization of the data

The first step in the data symbolization is the transformation of the original measurements into symbols. To do this, the state space is partitioned in a finite number of regions and then a symbol is assigned to each measurement depending on which region the observation did fall. In Figure 1 we can see a simple example of symbolization of an artificial time series of 20 signals: the alphabet has two symbols 0 and 1 and the binary partition is made using a frontier value *v*. If the original measurement is below *v*, we associate it with 0 and otherwise we symbolize it by 1.[5]

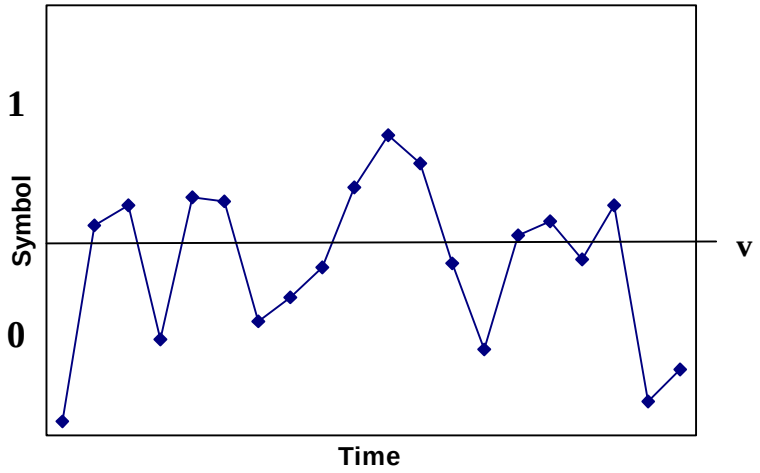


Figure 1: Illustration of symbolic encoding. The dashed line represents the symbol partition; data below the partition are represented by 0, and data above the partition are represented by 1. In this case, the frontier level has been chosen in such a way that the probability of each symbol is equal. Then, the original data time series is represented by the symbol sequence: S = 01101100011100110100.

If we have continuous data, we first have to choose an interval of time for the symbolization (i.e., we discretize the data) and this must be done with some ability so that we don't obtain excessive repetitions of the same symbol but taking care not to miss important frequencies.[6] In this case to obtain information about how the data change meaningfully in time, it is necessary to choose a symbolization interval which defines the number of actual data points between successive symbols. For the present work, we discuss only discrete time series. Although the choice of how to partition the data space affects directly the characteristics of the symbolic representation, the determination of the number of quantization pieces and the way we do it is ad hoc.

---

[4] Real data is always disturbed by noise, and in this case the concept of generating partition is usually not clearly defined. See J.P. Cruchfield and N.H. Packard (1983) for a detailed discussion.

[5] These kinds of symbolizations, where the phase space is separated into pieces labeled by symbols are usually called **static transformations**. We can distinguish them from the **dynamical transformations**, where the step-to-step-differences in the sequence are taken into account. A mixing of both kinds of transformations can be used. See J. Kurths, U. Schwarz, A. Witt, R. Th. Krampe and M. Abel (1996) for a detailed exposition. In our approach we will use only static transformations.

[6] See Daw, Finney, Nguyen and Halow (1998a) for an application of symbol statistics to continuously measured signals.

There is no general rule for selecting an optimal symbolizing partition although we know from symbolic dynamic theory that when we are in presence of some particular deterministic dynamical systems we can construct the so called Markov partitions where the symbolic dynamics is topologically conjugate to the original dynamical system. But in presence of any noise or for a large class of deterministic dynamical systems the identification of Markov partitions turns up to be impracticable. One criteria to symbolize the data used in Daw, Finney, Nguyen and Halow (1998b) is to define discretization partitions such that: 1) the occurrence frequency of any particular symbol is equiprobable to the others, or 2) the measurement range covered by each region is equal. This type of equiprobable partitions allows distinction between stochastic and deterministic structure in the data and so can be a contribution to the debate about the structure of business cycle time series. The partition shown in Figure 1 is defined using this criterion: the resulting time series has an equal number of zeros and ones.

In our context, the partition is selected by taking account of the different qualitative modes of behavior of the economy that we know a priori; that is, we divide the state space into *phase zones* according to different structures of the economy. This way of constructing partitions is helpful when we look for structural change in economic development and when we work with multiple regime models.[7] It is clear that the choice of the partition in the state space is a very important point in the analysis because the resulting symbolic sequence depends directly of the specific partition. Sometimes we can construct other partitions or refine the original partition in order to compare the obtained results.

3 Symbol sequence histograms

Once we have transformed the original measurements into a symbol string, we can begin with the characterization of the temporal patterns. In this direction, we introduce the *symbolic trees* and the *symbol sequence histograms*. The *symbolic tree* is a graphical representation of the symbol statistics as a function of the length of the words: we calculate the probability of occurrence for different words in the symbolic string and then we represent this in a tree. The first level of the tree are the probabilities of occurrence of the symbols, the second the probabilities of occurrence of the words of two symbols and so on. The symbolic tree is a compact summary of (coarse grained) information about the signal and each level of the tree corresponds to a particular length of symbol sequence. The following is an example of a tree for a binary symbolic sequence:

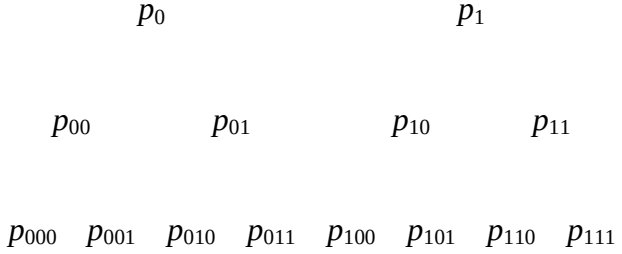


Figure 2: Symbolic tree representing the symbol sequence statistics.
Note that p_w is the probabilities of occurrence of the binary word w.

[7] This symbolization is exogenous because we construct the partition independently of the observed time series data. From the other hand, the symbolization shown in figure 1 is endogenous. Note that every dynamical transformation generates an endogenous symbolization.

To construct the *symbol sequence histogram*, we choose a symbol sequence length m and then we compute the relative frequency of occurrence of all possible symbol sequences of length m in the full symbol sequence. If we have a partition of n pieces (i.e., n symbols), we are to calculate the relative frequency of all the n^m words of m symbols that occur in the symbolic data sequence. We can describe this observed dynamics by a histogram of the relative frequencies. For easy reference and identification, a convenient representation of each word of length m is achieved by converting the base-n sequence into its decimal equivalent; for example, if $n = 2$ and $m = 4$ we will have $16 = 2^4$ different words of length 4 represented each one by a number from 0 to 15 and in particular the word 1101 is represented by the number $1.2^0+0.2^1+1.2^2+1.2^3 = 11$. This representation is called the *sequence code* and will help the computation of the histograms and visualising the results. In Figure 3 we show the histograms for the symbolic sequence of Figure 1 and $m = 2,3,4,5$.

Symbol sequence histograms are compact representations of the overall dynamics of time series and can be used to compare different sets of data. With these frequencies one can distinguish rather uniform distributions from more complicated ones. This leads to the first measure of complexity of the system which simply counts the "forbidden words" (or for the statistical approach, the number of trajectories with a low probability of occurrence). The selection of the sequence length m (like the selection of the partition) can be made in several ways and in part optimal choices depend on the way we have chosen the partition. It is clear that we can also construct different histograms for diverse values of m and compare them. Note that when we work with equiprobable partitions, symbol sequence histograms are useful for quantifying time irreversibility because the relative frequencies of certain symbol sequences will shift when the data are observed in reversed time. Conversely, for time reversible data series, there should be no significant difference in the histogram whether it is constructed in forward or reverse time.

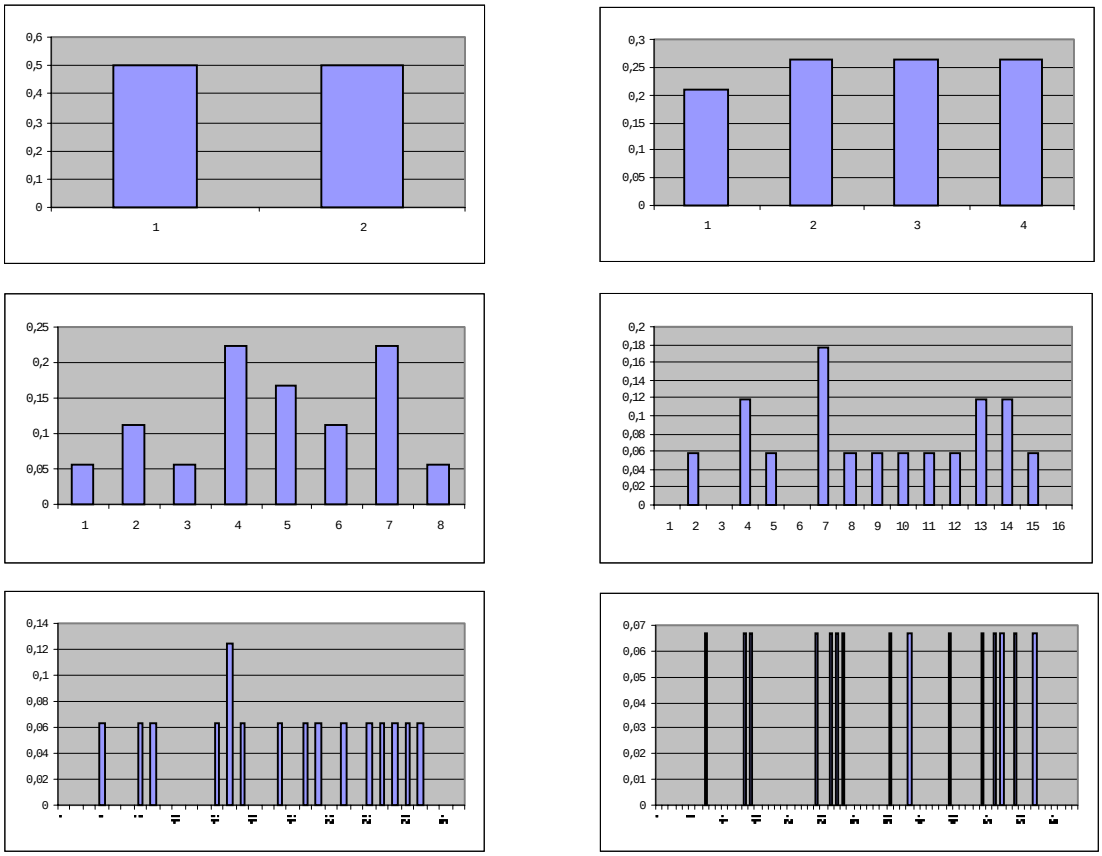


Figure 3: Histograms for $m = 1, 2, 3, 4, 5, 6$ for the symbolic time series of Figure 3.1.

The next step is the introduction of some statistical tools and information theoretical measures to characterize and compare different histograms:

1) The traditional quantity for characterizing a symbol sequence is the *Shannon entropy*.[8] The Shannon entropy of *m* order *H(m)* is based on the probability distribution of sequences with length *m* of the symbol sequence:

$$H(m) = -\sum_i p_i \log_2(p_i).$$

Here p_i is the probability of finding the *i*th sequence of length *m* (i.e., the number of times that the sequence indexed by *i* can be found in the long symbolic sequence divided by the number of all short sequences). This rather abstract quantity *H(m)* is simple to interpret: it measures the average number of bits needed to specify an arbitrary sequence of length *m* in a symbolic sequence. It is a measure of the complexity or variety of the process that generates the sequence. When we have a good partition, the symbolic signal will contain essentially all the information contained in the original signal. Note that Shannon entropy is a measure of *observational variety*: unobserved possibilities do not enter the measure and in the relative form can be interpreted as an index of uniformity (analogous to the standard deviation). The average is computed only for categories of observations that effectively occur and this is reflected by the convention $0.\log_2 0 = 0$. Further, we can consider the differences of Shannon entropy

$$h(m) = H(m+1) - H(m);$$

$$h(0) = H(1)$$

These differences quantifies the information needed to determine the $(m+1)$th symbol of an arbitrary sequence of a given symbolic sequence if the first *m* symbols are known. The Shannon entropy of the source is then defined as the limit of the *h(m)*:

$$h = \lim_{m \to \infty} h(m)$$

The *modified (or normalized) Shannon entropy* is defined as:

$$H_S(m) = -\frac{1}{\log_2 N} \sum_i p_i \log_2(p_i)$$

with $0.\log_2(0) = 0$ and where *N* is the total number of observed sequences with length *m* (i.e., the number of sequences of length *m* with non-zero frequency), *i* is a index for the sequences of length *m* and p_i is the probability of the *i*th word (i.e., the word designed by the index *i*). The choice of log = $\log_2$ reflects the fact that information is measured in bits. Clearly H_S is a normalized form of *H*

[8] The Shannon entropy, which is a natural measure of uncertainty and information for events characterized by probability distributions, has dominated the literature on information theory since it was proposed by Claude Shannon in 1948. It was originally introduced for the purpose of analyzing and designing telecommunication systems, but its significance and applicability reaches far beyond this original purpose. See Shannon (1963).

and can be interpreted as the average uncertainty per symbol.[9] A simple calculation shows that for any probability vector (p_1,…, p_k) of length k we have

$$0 \leq H_S \leq 1$$

with

$$H_S = 1 \text{ if and only if } (p_1,\ldots, p_k) = (1/k,\ldots, 1/k)$$

and

$$H_S = 0 \text{ if and only if } p_i = 1 \text{ for some } i.$$

The following is a list of some useful characteristics and interpretations of the different types of entropies.

i) Entropy is maximized when the probabilities are equally divided, and entropy is minimized when all the probability is concentrated in one component. In other words, entropy is zero when observational variety is absent and is maximum when all the categories are occupied by the same number of observations. In this sense, Shannon entropy is a *measure of randomness*.

ii) Entropies do not respond to the nature of the categories involved. Their labels are freely permutable: only the set of frequencies matter. It is in this sense that entropies are said to be *context-free*.

iii) Unlike variance measures of deviations from a mean which assume normal distributions, entropies assume nothing about the nature of the frequency or probability distribution they assess and are then non-parametric measures of variety and entirely general in this respect. Measures with this property are said to be *distribution-free*.

iv) Entropies are *averages*: one might interpret them as an average number of binary decisions made in the course of the classification.

v) Entropies are a function of relative magnitudes, probabilities being the most common form. The sample size does not influence the entropy values; it is for this reason that we have standardized them expressing the entropies relative to their maximum.

This type of statistics can be useful when we work with equiprobable partitions because for truly random data and with a sufficiently large set of data, each symbol sequence of length *m* must be equiprobable. Thus, a significative deviation from equiprobability is evidence of time dependence and deterministic structure in the data. For random data and equiprobable partitions we should have $H_S = 1$ and for non-random data it should be $0 \leq H_S < 1$ and lower H_S implies more deterministic structure.

The characterization of symbolic sequences is not restricted to the estimate of Shannon entropies. A broad range of so called measures of complexity allows a more detailed characterization of the structure of these symbolic sequences. One generalization of Shannon entropy is the *q*-Renyi entropy

[9] In Tang, Tracy and Brown (1997), *N* is taken to be the total number of possible sequences of length *m* but our election, according with Daw, Finney, Nguyen and Halow (1998), reflects the fact that many possible sequences may not be realized because of the finite length of the strings. The result is to bias H_S upward when the number of possible sequences became large to the available data.

$$H^{(q)}(m) = \frac{1}{1-q}\log_2\left(\sum_i (p_i)^q\right),$$

where q is a real number, $q \neq 1$. This measure converges to Shannon entropy as $q \to 1$. Due to different averaging it reflects some inhomogeneity of the underlying probability distribution. If $q > 1$ those sequences of length m with large probability dominate the Renyi entropy. For $0 < q < 1$ we have the reverse behavior.

2) The Euclidean norm is defined as

$$T_{AB}(m) = \sqrt{\sum_i (A_i - B_i)^2},$$

and the modified χ^2 is defined by

$$\chi^2_{AB}(m) = \sum_i \frac{(A_i - B_i)^2}{(A_i + B_i)},$$

where A_i and B_i are the individual sequence probabilities for the possible sequence codes i and for the histograms A and B. Both of these statistics can help when we want to compare different symbolic sequences.

The Euclidean norm works like a metric in the space of the possible sequences providing a magnitude of the distance between different histograms: a larger distance between histograms implies that the dynamics in the data set are very different. Also T_{AB} can be useful for testing for time reversibility on the time series data by comparing the histogram for a time series with that of its time reverse: the magnitude of the Euclidean norm between the histogram frequencies for the forward and reverse time quantifies the level of time irreversibility. The modified χ^2 is derived from the standard χ^2 by replacing the univariate measurement frequencies with sequence frequencies. Like the symbol sequence histograms, all these statistics are functions of the symbol sequence length m and there are no theoretical rules to determine which m is sufficient or the best for observing significant sequence patterns, if they exist. In Daw, Finney and Green (1998)**,** there is an empirical approach to select an appropriate m using the modified Shannon entropy. They show that $H_S(m)$ typically reaches a minimum value as m is increased from 1.[10] The explanation for this minimum in $H_S(m)$ is that reflects the symbol-sequence transformation which best distinguishes the data from a random sequence. Symbol sequences that are too short lose some of the important deterministic information and symbol sequences that are too long reflect noise and data depletion (i.e., there is not enough data to get reliable statistics for such long sequences). Then we can argument that the value of the symbol sequence length m that minimize the modified Shannon entropy is an "optimal" choice for the given data and election of the partition.

4 Symbolic analysis of economic time series data

In this section I report some applications of STSA to study the evolution and structural change of the Italian economy. We use data symbolization for the comparative analysis of the national-regional dynamics of accumulation, technical change and employment. We will base this section mainly on different works of Punzo described in section 2.4 of this thesis. We think that

[10] In Figure 7 we can see the graphs of $H_S(m)$ for the symbolization of our economic time series data.

data symbolization methods can be useful for characterizing and monitoring patterns in the context of these works, where a natural symbolization of the data can be built using economic arguments.

4.1 The Framework Space

The applications reviewed in this section manipulate time series of *value added* (VA), *gross physical capital formation* (I), both taken in real term, and *employment* (E). They are taken to reflect the impact of technological change and innovations, on which we focus on to explain the dynamics of the Italian economy and four macro-regions of Italy during the last two decades.[11] Growth rates are chosen as state variables and then only two co-ordinates were used in the model. The first co-ordinate is the rate of growth of value added per capita

$$(1) \qquad v = \frac{E}{VA}\frac{d}{dt}\left(\frac{VA}{E}\right)$$

while the other co-ordinate is the rate of growth of physical capital formation per person employed

$$(2) \qquad i = \frac{E}{I}\frac{d}{dt}\left(\frac{I}{E}\right).$$

A pair of corresponding values (v, i) will give coordinates in the plane R^2 which will be called the Framework Space (FS). The v-axe is called the Innovation axe and is and indicator of productivity growth; the i-axe called the Accumulation axe indicates the investment intensity. The former is associated with the Neo-Schumpeterian interpretation of productivity dynamics and the other with conventional theories of growth and technological progress. In the present framework, the two axes are plotted one against the other to obtain a unified analytical scheme and a framework to confront the two theoretical traditions. In this way we can also take account of interdependencies between the dynamic behaviors of I and VA. The dynamics of the economy is then be represented by the time-variation of the pairs (v, i) (the state variable) and the dynamical manifold of the system is R^2. For each one of the economies, there is a distinct Framework Space and so, to represent the evolution over a given time period, a convenient temporal segmentation is used. Once this has been chosen, a state represents the growth path of an economy at a given data. The innovation if comparing with the conventional approaches is that the path is observed now via two variables simultaneously.

It is a very difficult problem (if not impossible) to recover the underlying dynamical system in the two dimensional FS that generates the actual history of one of the economies. However, one such system does exist conceptually, and it might take either a continuous time formulation

$$(3) \qquad \dot{z} = \varphi(z)$$

or a discrete time formulation

$$(4) \qquad z_t = \phi(z_{t-1})$$

[11] The division of Italy in macro-regions is based in a geographical criteria and the regions are: North-Centre (NC), North-East (NE), West-Centre (WC) and Mezzogiorno (MZ). See Punzo (1996) for a description of the territories integrating these macro-regions.

where it is $z = (v, i)$. Both are two-dimensional systems of first order equations, the former a differential one and the other in differences. Note that we have a lot of problems with the mathematical formulation of the model: the election of time formulation, the identification of the explicit form of the system (i.e. the functions φ or ϕ), the scarce statistics, the dynamical interdependence among the economy in question with other economies, etc.

In order to recover the above dynamical systems, we will follow the heuristic approach adopted on the cited works of Punzo. If (3) or (4) describes the overall dynamical model of an economy, we can split it into a set of local models that are in principle simpler than the overall model.[12] So, we will have a twofold dynamics, one inside each local model and the other connecting the different local models. To recover the global model we can estimate first the local models and then the dynamics across the set of local models.
Local models are in principle independent one to each other, since they have to exhibit specific dynamical features. So, switches from one local model to another are qualitative changes (a change of "model") and then the dynamics across local models is a qualitative dynamics of the economic system. The analysis of the following sections will be particularly directed to recover the dynamics across local models rather than the dynamics inside each local model.

4.2 Regimes

Roughly speaking, a dynamic regime is a qualitative behavior that can be usefully distinguished from another dynamical behavior. Then, in the logic of the previous paragraph, we can identify dynamics regimes with local models. The formal definition of regime is the following: if (ϕ, R^2) is the dynamical system representing the economy, a division in regimes is a set of pairs $\{(\phi_1, S_1), (\phi_2, S_2),\ldots, (\phi_N, S_N)\}$, where $\{S_1, S_2,\ldots, S_N\}$ is a partition of R^2 and ϕ_i is the restriction of the dynamical rule ϕ to the set S_i $(i = 1,\ldots, N)$. Each pair (ϕ_S, S) is called regime.[13] We have a twofold dynamics, one within a given regime and one across regimes. Traditional dynamics focuses upon the former, overlooking the latter that will be the center of our approach. We will call it regime dynamics.

In this context, a regime switch is conceived as a qualitative change in the local model representing the regime. Then we can associate this qualitative change to a structural change in the history of the economic system under consideration. Therefore, regime dynamics describe structural changes in the economic system. If we consider the dynamics in the *FS*, a change of regime is always a movement from one path to another. A path across regimes is the equivalent of a Traverse in the Hicksian terminology.[14]

The ratios between growth rates in the *FS* give us one of the parameters of the *canonical model*. It can be either larger or smaller than one, and a ratio of exactly one can be treated as a bifurcation value. The set of all paths with investment and value added growing at the same rate (i.e. the 45° line) is called the Harrodian line having the typical property of Harrodian paths, its being a knife-edge. Ratio one corresponds to all paths belonging to the Harrodian line and it can be used to characterize that set compared to all others. On the other hand, the four semi-axes can be used to obtain the second parameter. All pairs of values of growth rates in the first and third quadrants preserve the same signs, while for paths in the 2 and 4 quadrants signs are interchanged.

[12] In principle the global system is non-linear while the local models could have a linear formulation. In this sense, the approach is close related to complex system theory.
[13] Observe that $\phi_i(S_i)$ is not necessarily a subset of S_i. Thus, paths can traverse from one regime to another.
[14] See Hicks (1973).

This reflects the fact that the underlying relationship between levels of variables, i.e. v and i, is increasing or decreasing. This can be represented by a second parameter ranging on the real line: for positive (or negative values) we get thus either relationship. The *Harrodian line* together with the other 4 semi-axes in a *FS* can now be used to induce a particular partition into *dynamical regimes*, each corresponding to a family of realizations of the *canonical model* for values of the two parameters in partitions of the parameter space induced by their bifurcation values, 1 and 0, respectively.

The Harrodian line divides the *vi* co-ordinate plane first quadrant into two regions where the economic system has different qualitative dynamics and then it can be represented by different local models. Paths taking place in the subset $\{(v,i)/\ i \geq v \geq 0\}$, where productivity falls behind investment growth, can be associated with conventional growth theories and paths taking place in the subset $\{(v,i)/\ v \geq i \geq 0\}$, where positive productivity growth rates exceeds positive investment growth rates, are better associated with innovational or Neo-Schumpeterian theories. Here we can identify two possible dynamic regimes of growth and development. Let's call them Conventional Growth regime (Regime VI) and Innovational regime (Regime I) respectively. Mirror images of these regimes are obtained in the third quadrant by the Harrodian line; that is, Regime III = $\{(v,i)/\ 0 \geq v \geq i\}$ and Regime IV = $\{(v,i)/\ 0 \geq i \geq v\}$. These are clearly regimes of negative growth (recession regimes).

In quadrants two and four, the state variables have opposite sign: an increase (decrease) of capital intensity is associated with a pace of decrease (increase) in productivity per employed. Then, we can identify two new dynamic regimes in these two sets with their respective local model. It is interesting to observe that these regimes do not receive any attention of theorists although of the empirical evidence of paths taking place there. Following a clockwise numeration, we will identify the regime associated with the fourth quadrant with number II and the other by number V. Regime II showing negative investment growth rates but positive productivity growth can be associated with "restructuring phases".

With the quadrants numbered clockwise, beginning with the Innovation Regime - and observing that the positive and the negative quadrants are further subdivided by the *Harrodian line*- a classification is obtained in six regimes.

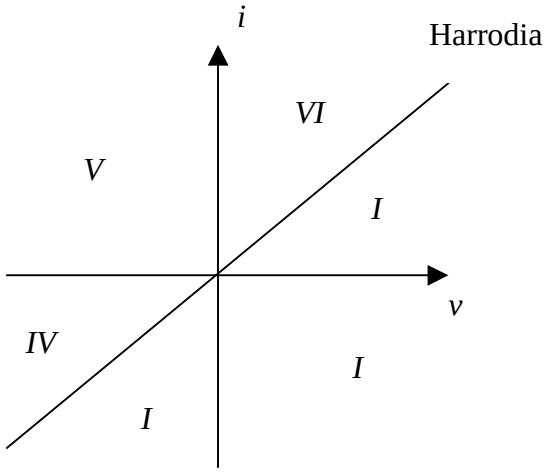


Figure 4: The Framework Space *iv* and the division in six dynamic regimes. This partition of the phase space give as a natural way to symbolize the data: each regime is labelled with its respective number.

It is only when the (*v*,*i*) co-ordinate plane is endowed with this induced partition, that it makes sense to call it *Framework Space*. Traditionally, theory sees only regimes VI and I (and their polar cases,

III and IV respectively)[15]. The introduction of regimes II and V gives us the possibility of analyzing oscillations that fall outside standard economic dynamics. Oscillations are now fundamentally across growth paths and these cannot be treated as (sometimes, purely virtual) long run equilibria or steady states. Dynamics that takes across regimes can be associated with structural change: it is the "model of growth" that is changing, not just its quantitative properties.

4.3 Data symbolization

In this application the data symbolization is based on economic criteria using the notion of *economic regime*. Once we have divided the state space of the dynamical system in regimes, we can do a statistical description of the underlying symbolic dynamics of the system. This dynamics is generated by the codification of each regime by one symbol and the translation of the punctual trajectory of the economy into a symbolic one. In order to highlight the proper features and relevant patterns in the symbolic series, we will follow an approximate approach to the six-pieces partition, starting with a two-pieces partition and refining it into three, four and six to arrive to the Framework Space with the division used in the cited works of Punzo. Each one of the successive partitions has, anyway, an economic interpretation. For example, the partition in two pieces separates a regime 0, in which both variables are positive of the other named 1, where one of them is negative. The partition in three pieces reflects a growth regime (1) with both *v* and *i* positive, one of reconstruction (2) where *v* is positive and *i* is negative and a third regime that includes the states in which the growth rate of gross investment is negative. We can state a similar interpretation for the partition into four pieces. This method of work, that for a moment moves us away from the work structure of Punzo, can be used to individualize with detail similarities and differences of the dynamics of the different economic systems and to compare them at different levels. It is for this reason that much of the statistic work will be repeated for each of the different partitions. See Figure 5 for a graphical representation of the successive partitions in dynamical regimes.

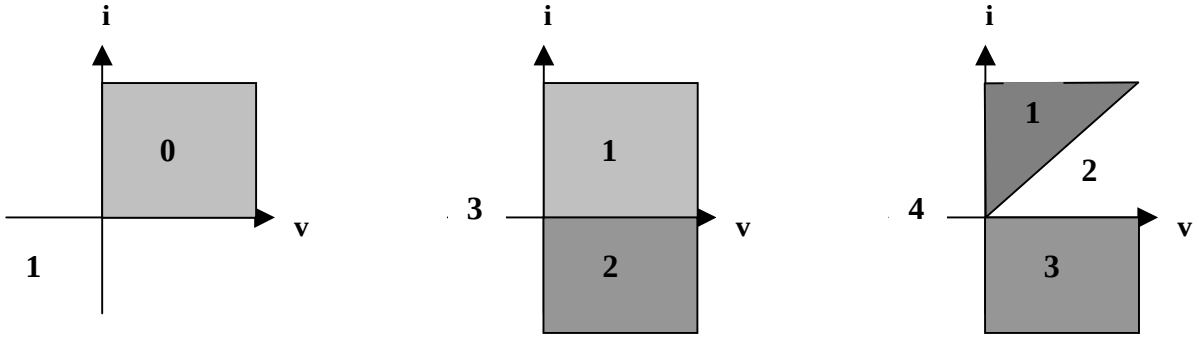


Figure 5: Partition of the Framework Space in 2, 3 and 4 dynamic regimes. Each partition is a refinement of the previous one.

We added an economic system to the five already existing. It correspond to the center of masses (barycentre) of the four Italian macro-economic regions and represents an average of them. The difference with the economic system Italy, that could also represent an average of the macro-regions, is that the barycentre is constructed from the symbolic trees and Italy is an aggregated of the macro-regions. Let $T_1(s)$, $T_2(s)$, $T_3(s)$ and $T_4(s)$ denote the symbol trees of MZ, NC, NE and WC respectively for the partition with s symbols, $s = 2, 3, 4, 6$. Let $T_{i,L}(s) \in R^N$ (with $N = s^L$) denote a

[15] Actually the former, i.e. the first quadrant, can be called the growth and the latter, the third quadrant, the contraction quadrants.

vector constructed from the Lth level of the ith symbol tree, i = 1, 2, 3, 4. This is a $N = s^L$-dimensional vector whose jth component is the probability of occurrence of the sequence with s-code j for the ith tree.

The center of mass of the four macro-regions, denoted as $T_B(s)$, is the tree whose Lth level is given by

$$T_{B,L}(s) = \frac{1}{4}\sum_{i=1}^{4} T_{i,L}(s).$$

The barycentre can be used like another average of the macro-regions (different from Italy) to measure the differences and similarities of the dynamical patterns.

Now we are ready to do comparative analysis of the different time series data, presenting them as symbolic sequences. If the fluctuations of two data series are governed by different dynamics, then the evolution of the symbolic sequences is not related. The symbolic sequences histograms of the 6 economies gives a reconstruction of their histories and provide a visual representation of the dynamic patterns. In addition, they may be used as basis to built statistics to compare regions that shows different dynamical properties. We can use symbol-sequence histograms to detect regime switches and to indicate which patterns are predominant.

4.4 Binary partition

In comparing the patterns of the economic systems, we begin with the binary partition evaluating the respective symbolic trees, the symbol-sequence histograms and the general trends of the modified Shannon entropy. Figure 6 illustrates the histograms produced using a binary partition and sequence length 6, Table 1 shows the most frequent words of each binary symbolic string and Figure 7 shows the graphs of the modified Shannon entropy as a function of sequence vector length.

Length	**ITALY**	**MZ**	**NC**
1	1	1	1
2	01 , 10 , 11	01 , 10 , 11	10 , 11
3	011 , 110 , 111	011 , 110 , 111	011 , 110 , 111
4	0111 , 1111	0111 , 1111	0111 , 1111
5	11011 , 11101 , 11111	11110, 11111	00111 , 11111
6	110111 , 111111	111110 , 111111	111101 , 000011
7	1111011 , 1111101 , 1111111	1111101 , 1111111	1001001, 1100100, 1101100, 1110110, 1111011, 1111011, 1111101, 1111110, 1111111
8	11111110 , 11111111	11111010 , 11111011, 11111101 , 11111111	11111110 , 11111111
9	111111110 , 111111111	111111110 , 111111111	111111101 , 111111111
10	1111110111 , 1111111011, 1111111110	1111101000 , 1111101111, 1111111010 , 1111111101, 1111111111	1011111111 , 111101100, 1111111010 , 111111101, 1111111110 , 1111111111
11	111111101111 , 11111111011, 11111111101	11111010001 , 11111111101,	11111111101 , 11111111111

		11111111111	
	NE	**WC**	**BAR**
1	1	1	1
2	10 , 11	10 , 11	10 , 11
3	101 , 111	011 , 110 , 111	011 , 111
4	1011, 1101 , 1111	0111 , 1110 , 1111	0111 , 1111
5	11011 , 11111	00111 , 11111	11110 , 11111
6	111110 , 111111	100111 , 111101 ,111111	111101 , 111111
7	1111101 , 1111111	1111011 , 1111111	1111101 , 1111111
8	11111101 , 11111110 , 1111111	11111110,11111111	11111110 , 11111111
9	111111101 , 111111111	101100111 , 110110011, 111011001 , 111111110, 111111111	111111101 , 111111110, 111111111
10	1101000011 , 1110100001, 1111110101 , 1111111011, 1111111101 , 1111111111	1101100111 , 1111111111	1111111101 , 1111111111
11	11101000011 , 11111111011, 11111111111	11101100111 , 11110110011, 11111111111	11111111101 , 11111111111

Table 1: Most frequent paths for Italy, the four Italian macro-regions and the barycentre with two regimes partition.

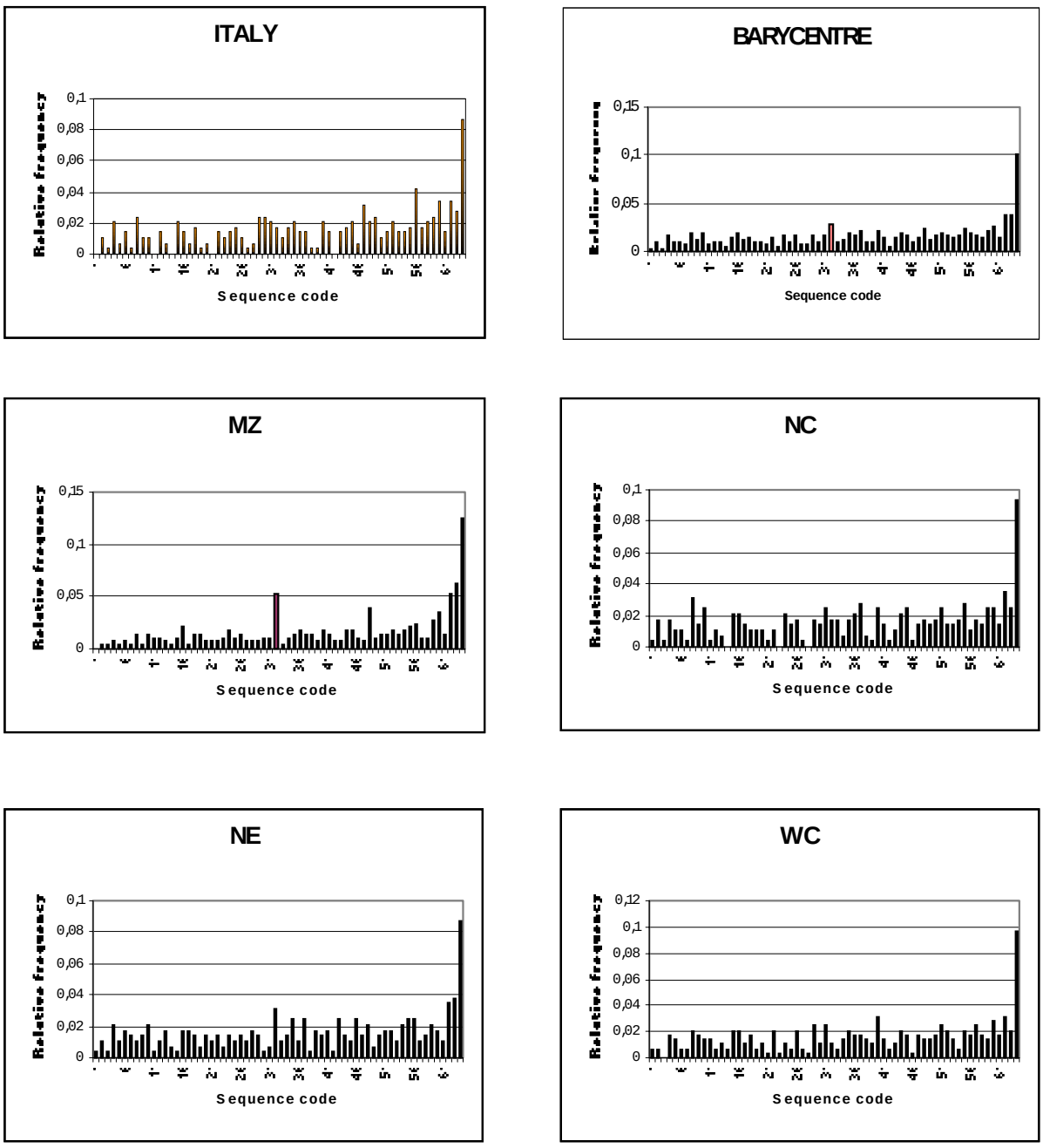


Figure 6: Symbol sequence histograms for Italy, the four regions and the barycentre of the regions. Two symbols of sequence length 6 were used. The dominant peaks on the histograms are often long strings of repeated ones.

These graphs show us that for all the economies the modified Shannon entropy reaches a minimum value at sequence vector length 6 or 7 as vector length is increased from one. We used the approximated optimal value 6 of the sequence length to do comparative analysis.

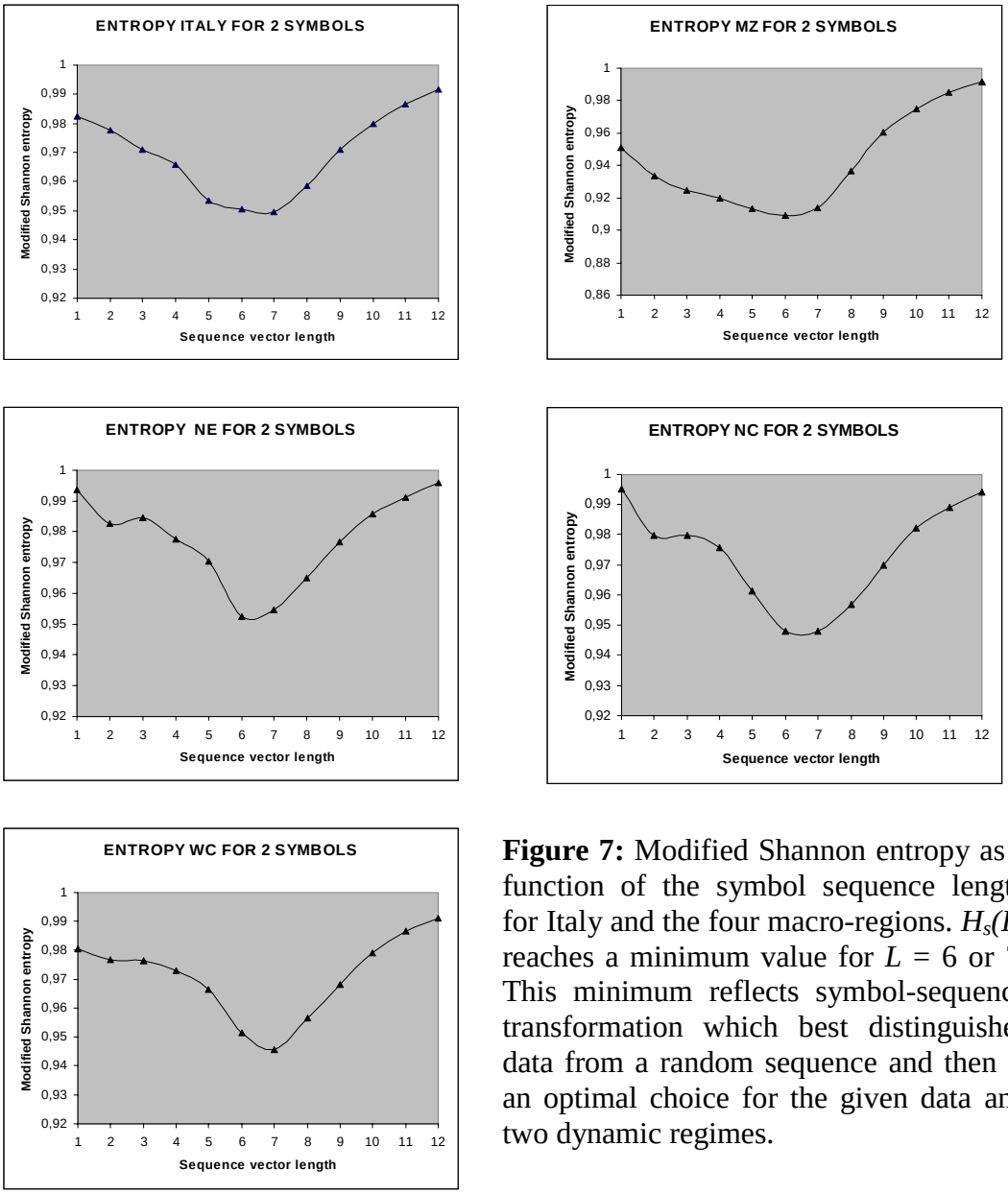


Figure 7: Modified Shannon entropy as a function of the symbol sequence length for Italy and the four macro-regions. $H_s(L)$ reaches a minimum value for L = 6 or 7. This minimum reflects symbol-sequence transformation which best distinguishes data from a random sequence and then is an optimal choice for the given data and two dynamic regimes.

For all the economic systems, regime 1 is hardly a little more probable than regime 0 (the probabilities of visiting 0 and 1 are in average of the order of $p_0 = 0{,}42$ and $p_1 = 0{,}58$) but observing Table 1 we can note that most frequent paths contains large strings of ones and, when a zero is present, then almost always the following symbol is 1. This fact reflects certain structural stability of this regime: once the economy enters regime 1 remains there for long time with high probability. And reflects certain instability of regime 0 or the improbability of periodic fluctuations from one regime to the other. So, there is some kind of "trap" at regime 1 where the economies live for long periods and we do not observe long periods in regime 0 but fleeting visits. Only NC shows some recurrent periods (three or four years) of relative stability at regime 0. At the optimal levels of the trees similar patterns for all the systems are visible. At level 6 there is a high frequency of symbol sequences 100111, 110111, 111101, 111110 and 111111; at level 7, symbol sequences 1111011, 1111101 and 1111111 are the most frequent. These sequences appear to occur because of transient or nonstationary dynamics in the systems. MZ is the system where the Shannon entropy takes lowest values at all levels of the tree; this can be interpreted like a minor complexity of its dynamics. Note that this can be the consequence of some kind of "isolation" in the dynamics of this

macro-region. It is clear that we are working with an economic system (Italy) divided into four sub-economic systems and we cannot isolate the dynamics of each macro-regions. There must be a strong relationship between the dynamics behavior of the regions and the dynamics The exercise of dividing the whole economic system into four sub-systems can be seen in the optics of complex system dynamics where the whole dynamics is made from each particular macro-region dynamics but fundamentally from the relationships between the regions.

The repetition of ones in the most frequent words gives reason to divide regime 1 into two pieces obtaining the three pieces partition. It is also clear that the adopted division tries to unfold the dynamics of the processes into a restructuring regime (named 2) and a low performance regime (named 3).

4.5 Three pieces partition

Figure 8 illustrates the histograms produced with the three pieces partition and sequence length 4, Table 2 shows the most frequent words of each symbolic string and Figure 9 shows the graphs of the modified Shannon entropy as a function of sequence vector length. Entropy is minimized at level four and we explain this minimum in H_S as reflecting the symbol sequence transformation which best distinguishes the data from a random sequence. Sequences vectors that are too short lose some of the important deterministic information. Sequence vectors that are too long reflect noise and data depletion (that is, there are not enough data to get reliable statistics for such long sequences). Thus, level four of the tree will be the target to do symbol sequence comparisons. The dominant peaks on the symbol sequence histograms of Italy are long strings of the same repeated symbols 1 or 2. This fact can be interpreted as some kind of stability on regimes 1 and 2. Note that regime 2 is also stable in the other systems but Italy is the unique system showing stability of regime 1. Another characteristic of Italy is that very few passages at regime 3 are observed for this system (only at levels 7 and 8).

Periodic strings of the symbols 1 and 2 are the dominant peaks of the symbol sequence histograms of MZ and then we can say that this system shows periodic (and sometimes not periodic) fluctuations between regimes 1 and 2. Frequent paths visiting regime 3 are observed at higher levels of the MZ tree and most of them are traverses from or to regime 1. Note that at level 8, MZ has only five frequent paths, two of them periodic fluctuations between regimes 1 and 2. At this level, the other systems have a “more equally” distribution showing nine or more frequent paths with no significant regularity.

NC has often passages from 1 to 3 and very frequent paths (of all levels of the tree) traversing these regimes, some of them almost periodic. There is no signal of stabilization of NC at any regime, but once the system enters regime 2 it seems to stay there rather than escaping to other regime. At level five and eight of the tree this system shows several dominant peaks on the histograms.

At the optimal level (four) of the tree, NE has the particularity of showing a (quasi) equally distribution (see Table 2 and figure 8) whereas the other systems have a peak at symbolic sequence 2222.[16] In fact, Ne is the unique system that has not frequent periods staying in a particular regime, showing more switches of regime than the other systems. Note that at the sixth level of the tree, NE has a bimodal distribution with peaks in two paths visiting regimes one and three with no particular regularity.

[16] Represented in the histograms of Figure 8 by the sequence code 41.

Length	ITALY	MZ	NC
1	1,2	1,2	1,2
2	11 , 22	21 , 22	11 , 13
3	111 , 222	212 , 222	131 , 222
4	1111 , 2222	1212, 2121, 2222	2131 , 2222
5	11111, 22222	22221 , 22222	11131, 11311 , 21312 , 21313 , 22221 , 22222
6	111111 , 222221	121212 , 131131 , 212121 , 222212 , 222221	111311 , 222221 , 222222
7	1111111, 1112213, 2131233, 2222211	1311313, 2121212, 2222212	1113113, 1131133, 2131233
8	11111111,1131333, 13133211,2131233,21313321, 22131233, 22222112, 32222231,33222223	11112311,1121231, 12121212, 21212121,2222212	11131133,1113311, 11311331, 13111131,13133211,13311131 , 21222222,21313321,22131233 , 31111311,31113113,33111311
	NE	**WC**	**BAR**
1	1,2	1,2	1,2
2	11 , 21	11, 22	11 , 22
3	111 , 121	131 , 222	131 , 222
4	1133 , 1212 , 1332 , 2222 , 3121	2222 , 3211	1131 , 2222
5	13133 , 13322 , 13331 , 33311	21312 , 22222, 33211	22221 , 22222
6	131133 , 133311	111221,112213,131233,133211, 133313,211311,213123,222221, 222222,312131,321131,331311, 332111,333131	222221 , 222222
7	1113133, 1121112, 1133311, 1211331, 1212222, 1311332, 1311333, 1313322, 1332213, 1333112, 2121232, 2122222, 2131133,2131332, 2212123, 2221212, 2221212, 2222132, 2323221, 3113331, 3133221, 3322131,3331121	1112213, 1333131, 2131233, 3211311	1222222, 2122222, 2131233, 2131332, 2222212, 2222221
8	12122222, 13111333, 13133221, 13322131, 13331121, 21313322, 22121332, 22212123, 31332213	11122133,11133111,11221333 , 11331113,12321131,13133211 , 13311131,13331311,21222222 , 21312333,22131233	11131133,11133111,11331113 , 12122222,13111131,13133211 , 13311131,21222222,22131233 , 22222112,31113113

Table 2: Most frequent paths for Italy, the four Italian macro-regions and the barycentre with three regimes partition. At the first three levels we have bimodal distributions with diverse peaks.

NC, WC and the barycentre have frequent paths symbolized by 131 at the third level of the tree. Switches on regimes 1 and 3 are also observed for these systems at higher levels, especially at levels 7 and 8.

Italy and the barycentre have a bimodal distribution at the first six level of the tree while the other systems have bimodal distribution only for the firsts, third or fourth levels. So, we can say that path dependence is stronger for the averages than from the other systems.

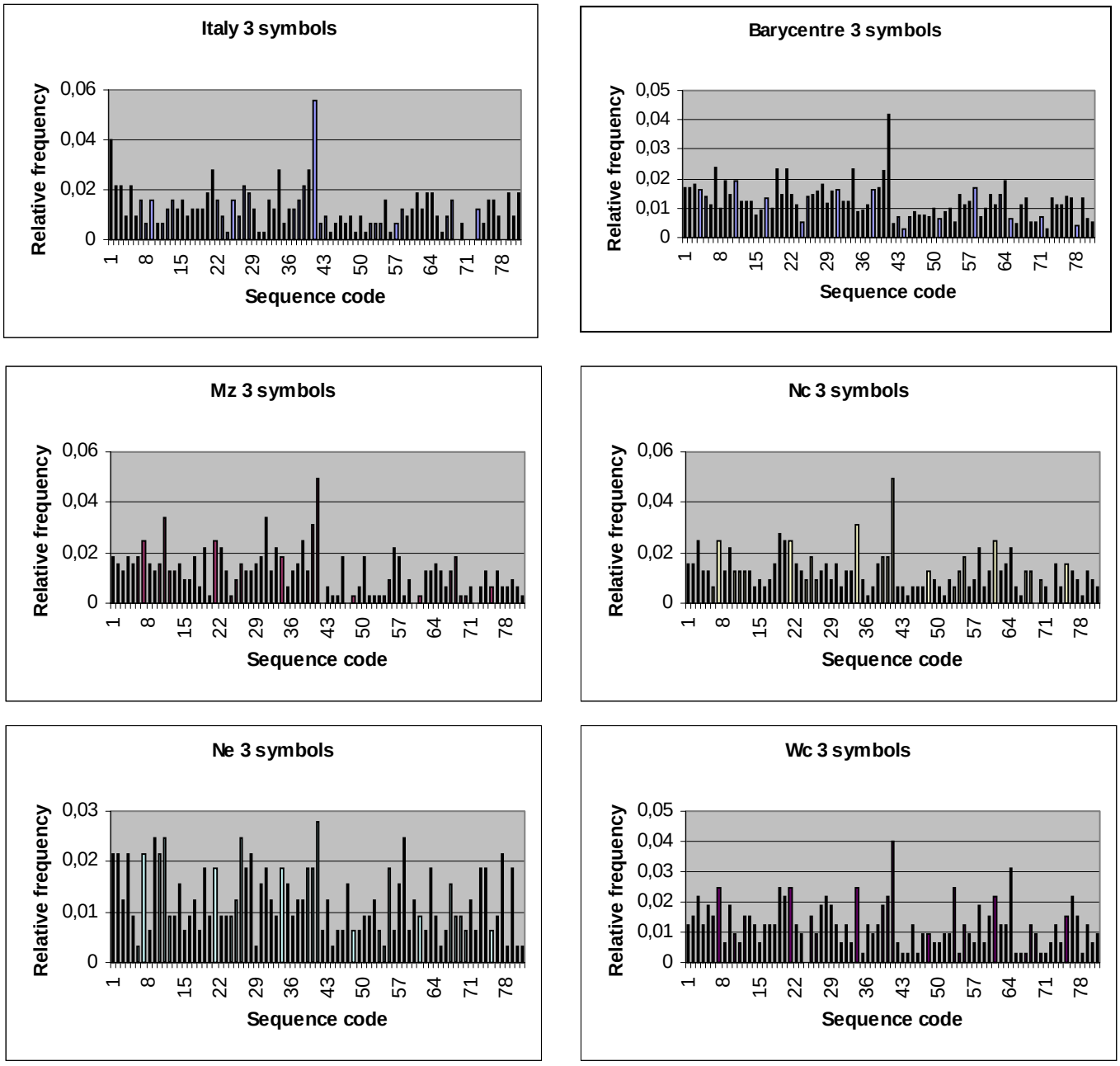
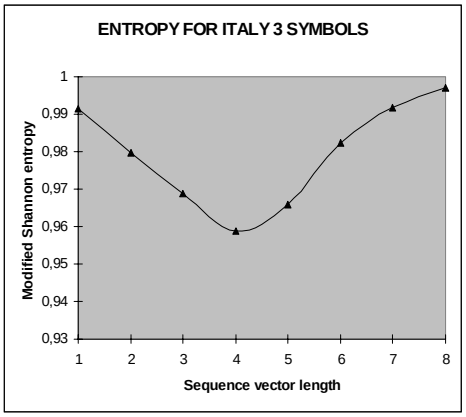
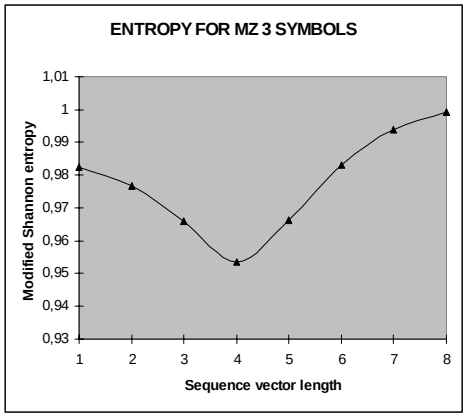
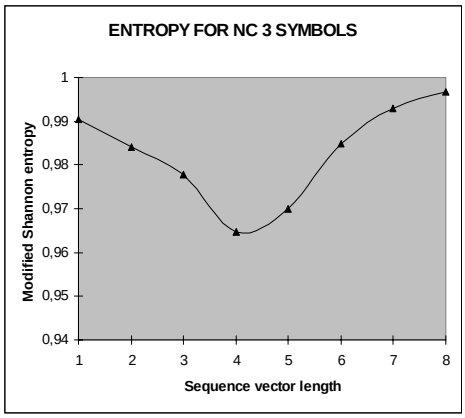
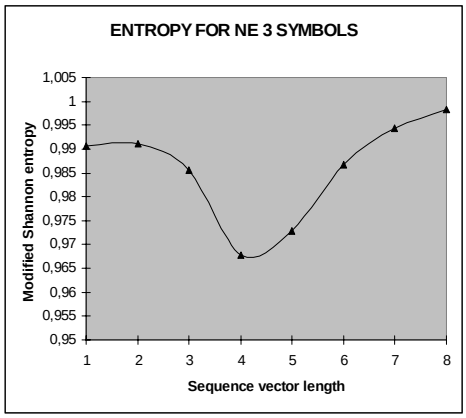


Figure 8: Symbol sequence histograms. Three symbols of sequence length 4 were used.









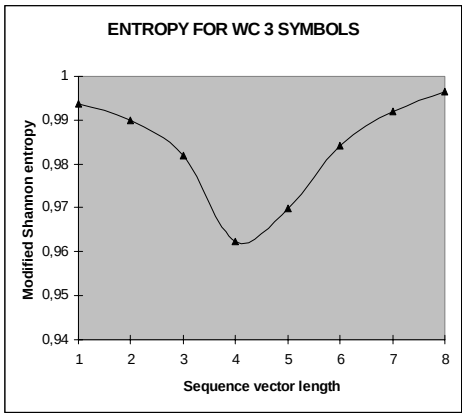


Figure 9: Modified Shannon entropy as a function of symbol sequence length for three regimes data symbolization. $H_S(L)$ reaches the minimum value for $L = 4$ and then this is the optimal value for the statistical analysis of the symbolic sequences of three symbols.

At this point, we divide regime 1 of the three pieces partition into two pieces, obtaining the four pieces partition. In this case, regimes 1 and 2 are respectively the Conventional Growth and Innovation regimes of Punzo's Framework Space.

4.6 Four pieces partition

Figure 10 illustrates the histograms produced with the four pieces partition and sequence length 3, Table 3 shows the most frequent words of each symbolic string and Figure 11 shows the

graphs of the modified Shannon entropy as a function of sequence vector length. Entropy is minimized for all the systems at level three; then we will use this level of the tree to do comparative analysis.

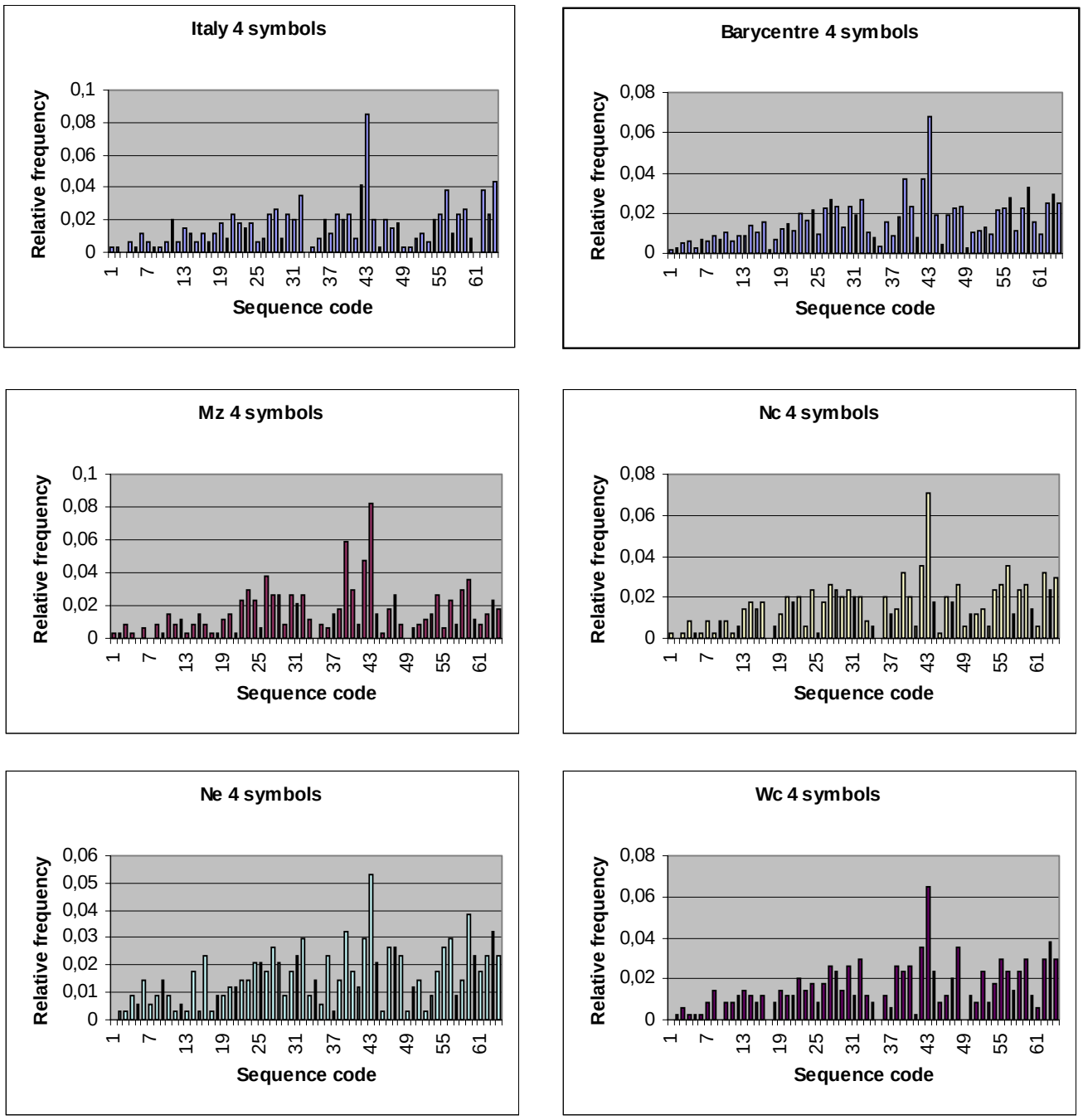


Figure 10: Symbol sequence histograms. Four symbols and sequence length 3 were used.

All the systems have frequent paths composed by large periods in regime 3.[17] Moreover, these paths are the most frequent at all. This fact confirms the structural stability of this regime stated at the three pieces partition. Most frequent traverses to (from) regime 3 are from (to) regime 2. Although regime 4 is frequently visited by MZ, there seems to be no regularity in the traverses to

[17] This can be observed in Figure 10, where the larger peak is at sequence code 43 (representing the symbolic sequence 333).

or from this regime. This is reflected by the fact that most frequent paths in MZ do not visit regime 4. Same thing can be stated to the barycentre.

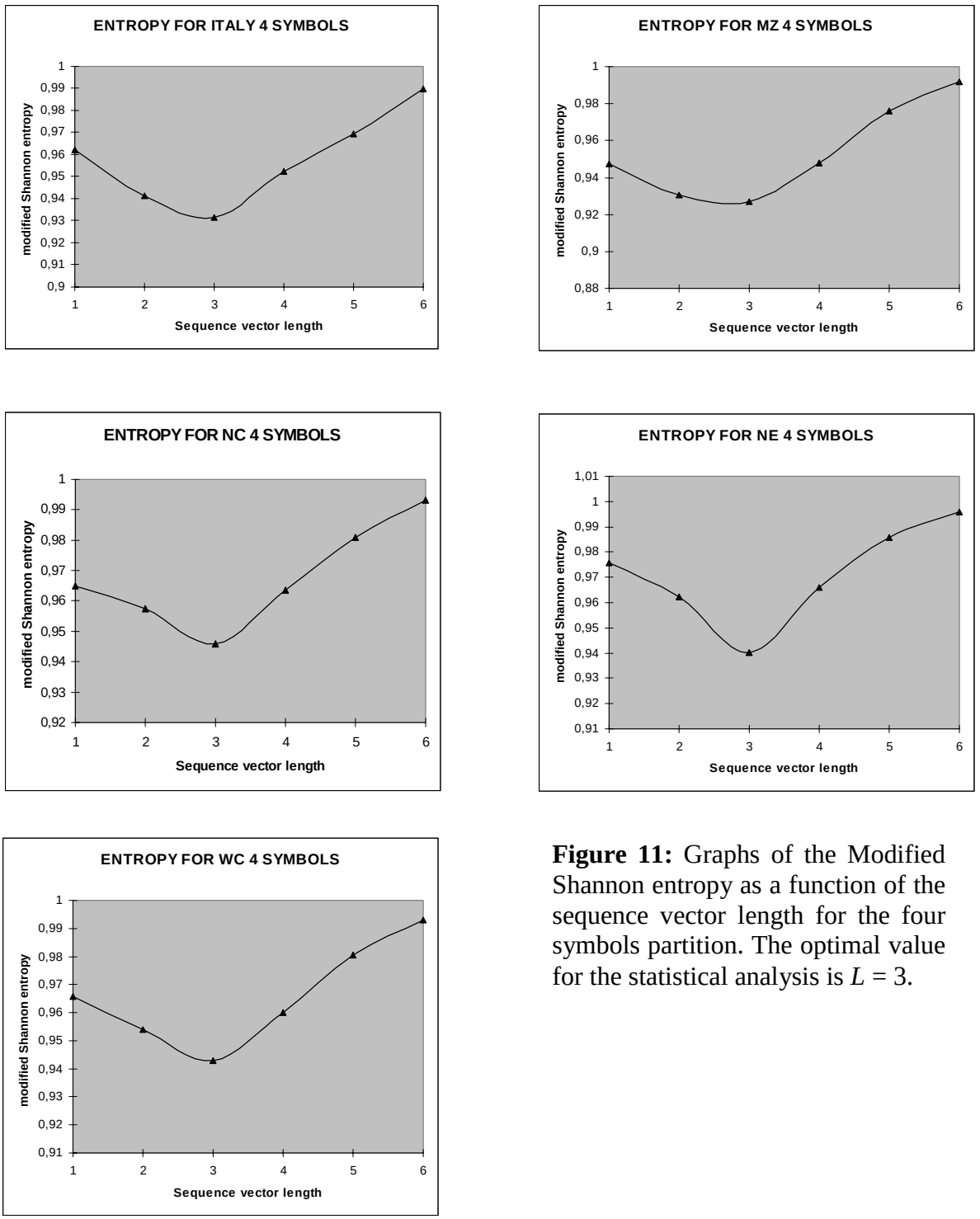


Figure 11: Graphs of the Modified Shannon entropy as a function of the sequence vector length for the four symbols partition. The optimal value for the statistical analysis is $L = 3$.

Italy has a bimodal distribution (with a large peak in symbol sequences with symbol 3) at all levels of the tree. Same thing can be stated to NC and the barycentre, with the exception that for these systems at level 3 there is a three-modal distribution. NE is characterized by a sixth level distribution with twenty peaks equally distributed and where there are frequent paths visiting regime 1. Is the unique system having frequent paths traversing regime 1 but this is not very

significant; it can be better interpreted like some random behavior at this level of the tree: at levels 1 to 5, the dynamical frequent paths of NC are not very different from the ones of the other systems.

Length	ITALY	MZ	NC
1	2,3,4	2,3,4	2,3,4
2	33,44	32,33	33,42
3	333,444	323,333	332,333,424
4	3332,3333	3232,3332,3333	3332,3333
5	33332,33333	33332,33333	33332,33333
6	333332,433333	232323,323232,333323,33333, 333333	333332,333333
	NE	**WC**	**BAR**
1	2,3,4	2,3,4	2,3,4
2	33,44	33,44	32,33
3	333,433	333,443	323,332,333
4	3333,4433	3333,3443	3332,3333
5	24433,33332,33333	33332,33333,44322	33332,33333
6	131144,141443,214431,224442, 232333,232343,233144,233333, 314244,323234,323333,332323, 333232,333243,333324,333332, 343433,424433,434332	242344,324234,333332,333333	333332,333333

Table 3: Most frequent paths for Italy, the four Italian macro-regions and the barycentre with four regimes partition.

4.7 Six pieces partition

Finally, we arrived to the partition in six regimes originally presented in the mentioned works of Punzo. Here the most important regimes (in the viewpoint of economic theory) are 6, 1 and 2, while regimes 3, 4 and 5 are their mirror images. Figure 12 illustrates the histograms produced with the six pieces partition and sequence length 2, Table 4 shows the most frequent words of each symbolic string and Figure 13 shows the graphs of the modified Shannon entropy as a function of sequence vector length.

Length	ITALY	MZ	NC
1	2 , 6	2 , 6	2 , 6
2	62,66	22,26	62,66
3	626,662,666	252,262,266,622	626,662
4	2666,6662,6666	2222,2525,2666,5252	1662,2216,3256,5626,6626
5	16266,26632,26666,26632, 62366,62666,66322,66626	25252,52525,62663	26236,62216,62366,66322
	NE	**WC**	**BAR**
1	2 , 6	2 , 6	2 , 6
2	62,66	26,62,66	26,62,66
3	266,626	266,366,662,666	266,626
4	2666,6266	2666,3665	2666,6266,6626
5	16266,66622	62366,66262,66322	26666,62366,62663, 66262,66322

Table 4: Most frequent paths for Italy, the four Italian macro-regions and the barycentre with six regimes partition.

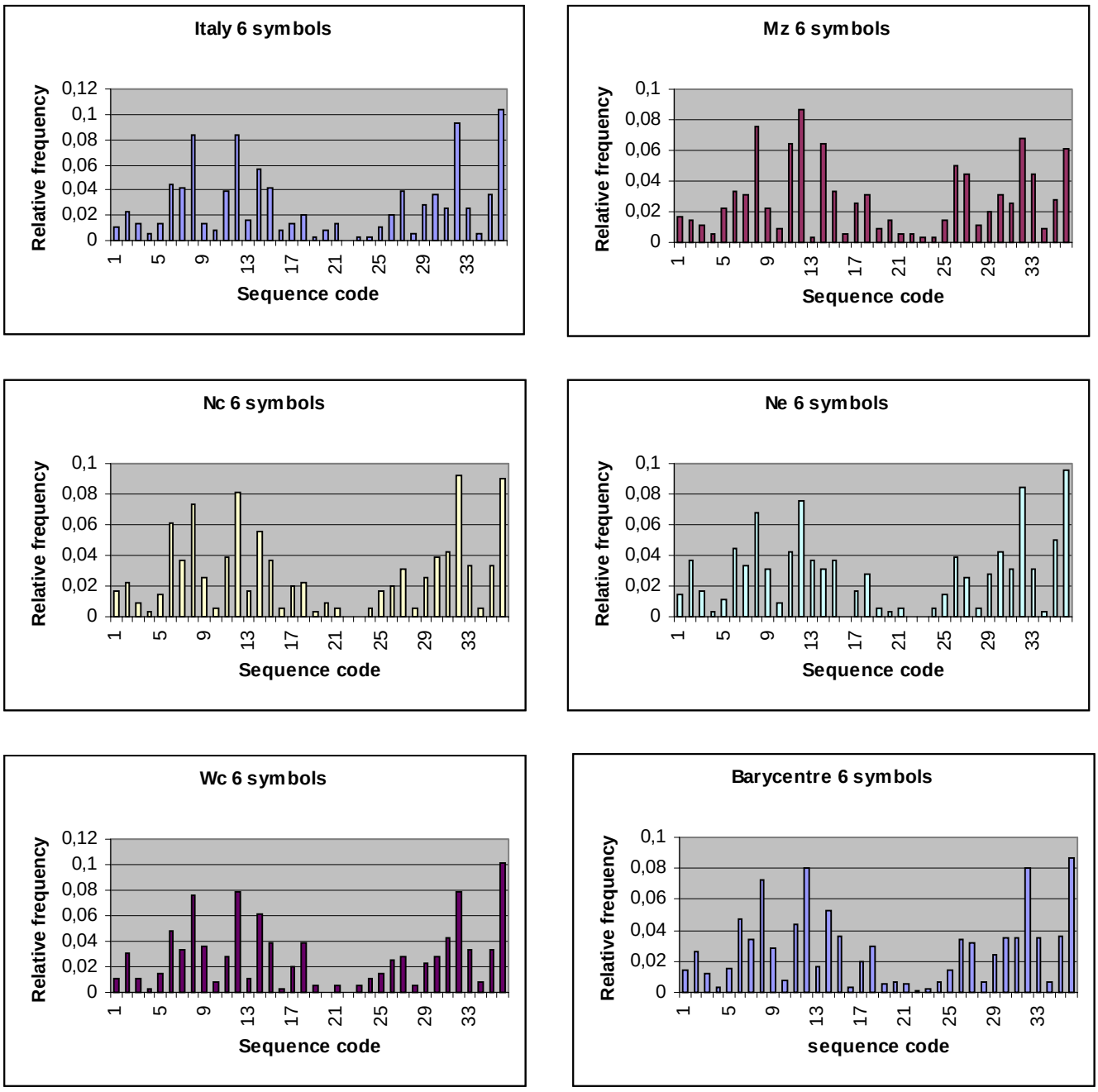


Figure 12: Symbol sequence histograms. Six symbols of sequence length 2 were used.

When we observe the first level of the trees, the first impression is the existence of a bimodal distribution for all the economic systems, with regimes 2 and 6 having the highest probability of being visited. In fact, if we assume that the process can be interpreted with a Markov chain, all the transition matrices satisfy the conditions for the existence of an ergodic distribution for each system. These distributions are again bimodal in all cases with regimes 2 and 6 being the most visited confirming the impression of the observation at the first level of the trees. At second and third level of the trees, most frequent paths of the economic systems visit regimes 2 and 6 (with

the exception of MZ and WC, where we can see the paths 252 and 366 respectively) but generally in an alternating way reflecting traverse paths to be more frequent than "stability" paths (i.e., paths visiting the same regime). For example, at the third tree level, only for Italy and WC we have a frequent stability path: 666. If we observe higher tree levels, it is more evident that there is very less stability in the regime dynamics and that oscillations paths visiting regimes 2 and 6 but not only are the most frequent. Note that MZ dynamics is quite different from the others showing frequent traverse paths visiting regimes 2 and 5 in a periodic way (2525, 5252, 25252, 52525). Nevertheless, the other economies show frequent trajectories visiting almost only regimes 2 and 6 with not evident regularity. It also seems that there is no convergence of the Italian macro-regions to a common dynamical path because the most frequent paths of the economies are substantially different.

NE is the unique system showing a bimodal distribution at all levels of the tree. Then, fluctuations seem to be the typical way in which the evolution of this economy take place, changing structure rather than stabilizing the process in a particular regime.

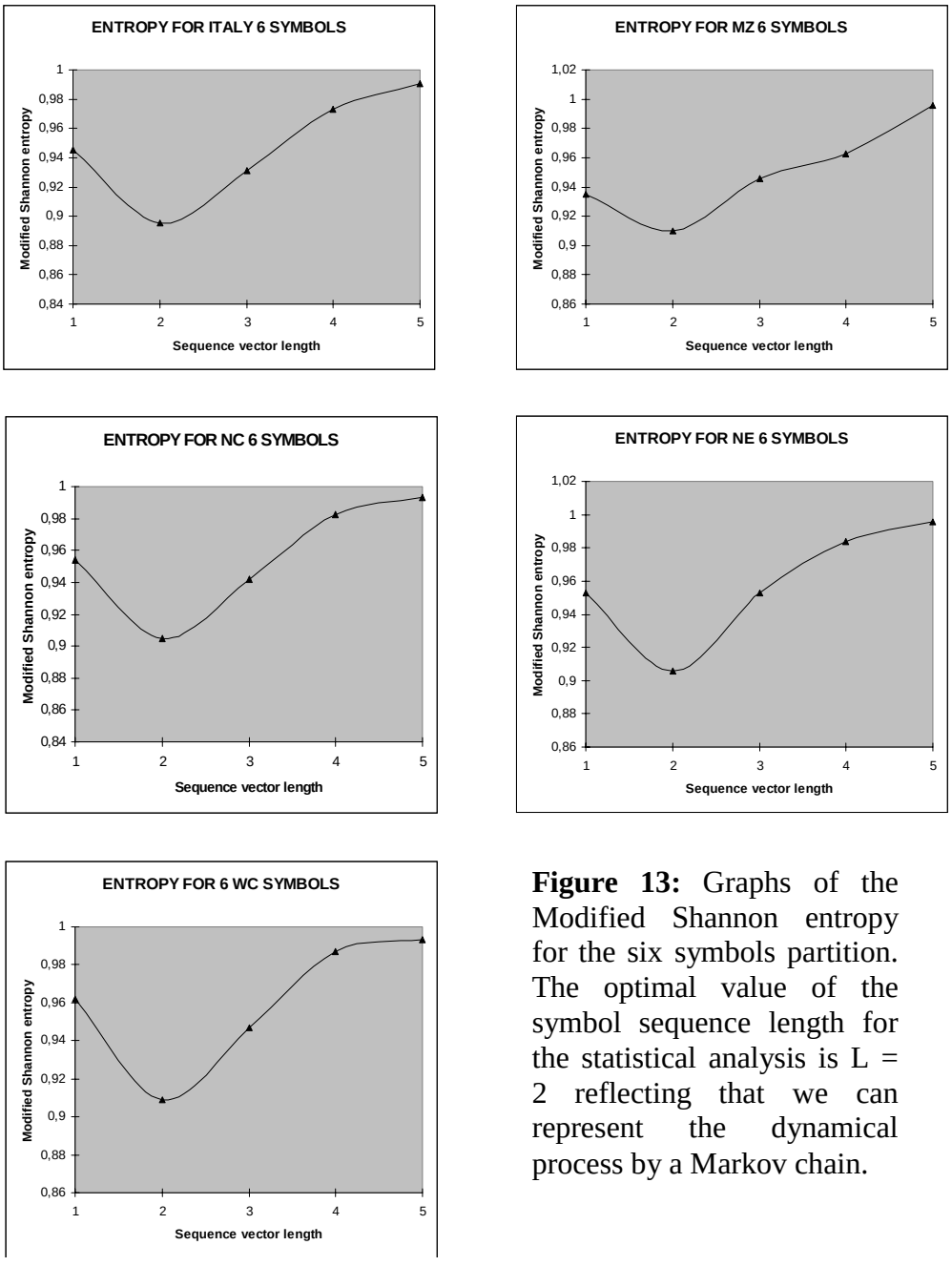


Figure 13: Graphs of the Modified Shannon entropy for the six symbols partition. The optimal value of the symbol sequence length for the statistical analysis is L = 2 reflecting that we can represent the dynamical process by a Markov chain.

Classical growth theories predict dynamical states belonging to regime 6 while theories where technical progress is the result of innovational activities focus on regime 1; both of them ignore the possibility of economies belonging to the other regimes. But it is clear from our analysis that, at least regimes 2 and 5 must not be disregarded. In particular, we saw that MZ seems to be well reflected by a periodic fluctuation between regimes 2 and 5 and the other economies seem to fluctuate with no regularity regimes 2 and 6. Very few frequent paths show passages at regime 1, and when this happens the next place in the sequence is regime 6.

4.8 Conclusions of the comparative analysis

From the analysis at the different refinements of the binary partition there are some recurrent facts and some differences that can be stated. MZ is the system that has more different dynamical behavior from Italy for all symbolization. This fact is also reflected in Table 5 showing the distance between the "optimal" levels of the trees.[18] NC has the nearest trees to Italy reflecting the fact that the regime dynamics of this macro-region is the most similar to the one of the whole country. Similar conclusions about the regime dynamics can be stated when we compare the barycentre with the macro-regions: WC is the most different and NC the most similar.[19]

	IT-BAR	IT-MZ	IT-NC	IT-NE	IT-WC
2 symbols - L = 7	0,050258	0,082208	0,057663	0,071668	0,062824
3 symbols - L = 4	0,050837	0,079054	0,051248	0,075073	0,055728
4 symbols - L = 3	0,053692	0,082667	0,051617	0,072523	0,057934
6 symbols - L = 2	0,041275	0,077222	0,038611	0,057406	0,049794

Table 5: Symbol statistics for comparisons of Italy with the macro-regions. The Euclidean norm T_{AB} for the difference of the two optimal symbol sequence histogram levels was computed for each symbolization. The distance between levels of the trees can be used as a diversity measure of the dynamics of two systems.

	BAR-MZ	BAR-NC	BAR-NE	BAR-WC
2 symbols - L = 7	0,053325	0,039404	0,047207	0,049376
3 symbols - L = 4	0,051911	0,034838	0,042067	0,040024
4 symbols - L = 3	0,052428	0,030742	0,040408	0,035447
6 symbols - L = 2	0,054144	0,028583	0,040459	0,032076

[18] Note that we stated our conclusions using only the "optimal" levels of the trees but these are true at almost all the levels of all the symbolic trees.
[19] See Table 6.

Table 6: Symbol statistics for comparisons of the barycentre with the macro-regions.

	MZ-NC	MZ-NE	MZ-WC	NC-NE	NC-WC	NE-WC
2 symbols - L = 7	0,080715	0,074986	0,091099	0,07279	0,058824	0,083997
3 symbols - L = 4	0,076840	0,073786	0,081912	0,06339	0,046336	0,069504
4 symbols - L = 3	0,072403	0,080655	0,076357	0,05455	0,046318	0,060276
6 symbols - L = 2	0,072829	0,083941	0,076096	0,05270	0,040008	0,056022

Table 7: Symbol statistics for comparisons of the macro-regions.

Although the construction of the symbolic trees of Italy and the barycentre is different, both can be seen as an average of the macro-regions. Remember that Italy trees were constructed from the symbolic sequence[20] while the barycentre trees are an average of the macro-regions trees. Note that this fact make not significant differences on their regime dynamics behavior.

In comparing the macro-regions, we can note that NC and WC have the closer dynamic behavior while MZ seems to have a different dynamics than the other macro-regions. Table 7 represents the Euclidean distances between the macro-regions at the optimal levels of the trees.
We want to emphasize that our estimates of symbol-sequence histograms will be uncertain to some extent because we use finite data sample for their construction. Thus, at some point it is important to establish confidence limits for our statistics. That is, how do we asses when the differences between estimated symbol-sequence histograms are large enough to warrant saying that the data sets being compared are generated by different dynamical systems within some specified level of confidence? This issue seems to be complicated because of strong relations in the dynamical behaviors of the systems making difficult to assume the hypothesis of independence used to establish confidence limits for standard statistical tests.

4.9 Time irreversibility analysis

Time irreversibility is defined such that a qualitative or quantitative description of a time series and its time-reversed realization differ significantly. Linear Gaussian random processes are inherently time reversible, whereas noisy non-linear dynamics systems are inherently time irreversible. While time irreversibility is not an absolute test for nonlinearity, its presence is a significant indicator of nonlinear structure. In our approach, we employ symbolic time-series analysis for quantitative evaluations of time irreversibility. Note that our successive partitions are (almost) equiprobable implying that the relative frequency for each symbolic sequence for truly random data will be equal (subject to the availability of sufficient data). Then any significant deviation from equiprobable sequences is indicative of time correlation and deterministic structure.

We can characterize the time irreversibility in each economic system by observing the optimal symbol-sequence histograms level for the forward-time and backward-time realizations. These differences can be quantified by the time irreversibility index T_{irr} that measures the Euclidean

[20] The symbolic sequence of Italy is constructed from the time series data, and this data is the aggregated of the macro-regions data.

distance of the forward-time and backward-time realizations. Then, the magnitude T_{irr} quantifies the level of time irreversibility. Recall that T_{irr} can take values between 0 and $\sqrt{2}$. Results for our time series are summarized in Table 8.

	ITALY	MZ	NC	NE	WC
2 symbols - L = 7	0,062608	0,058362	0,077468	0,064731	0,085592
3 symbols - L = 4	0,051806	0,074818	0,051806	0,059391	0,052540
4 symbols - L = 3	0,052941	0,072761	0,049913	0,046875	0,033792
6 symbols - L = 2	0,086336	0,084499	0,069860	0,043214	0,042665

Table 8: Time reversibility statistics for Italy and the macro-regions. Time irreversibility is defined such that a qualitative or quantitative description of a time series and its time-reversed realization differ significantly. We employ symbolic time-series analysis for quantitative evaluations of time irreversibility.

Time reversibility of our time series seems to depend on the size of the symbolization. For MZ the index T_{irr} increases with the size of the partition while for NC, NE and WC decreases. The maximal degree of time irreversibility at six pieces partition is seen for Italy and MZ.

Figure 14 illustrates symbol sequence histograms for the forward and reverse-time realizations of Italy data. The reverse-time realizations were produced by reversing the order of the time series. The larger differences are at sequence codes 9 (12) and 17 (27); i.e. at sequences 23 (32) and 25 (52) showing that there is some preference for the order in which they occur. This preference for order should have important implications for selecting the optimal model of the economy.

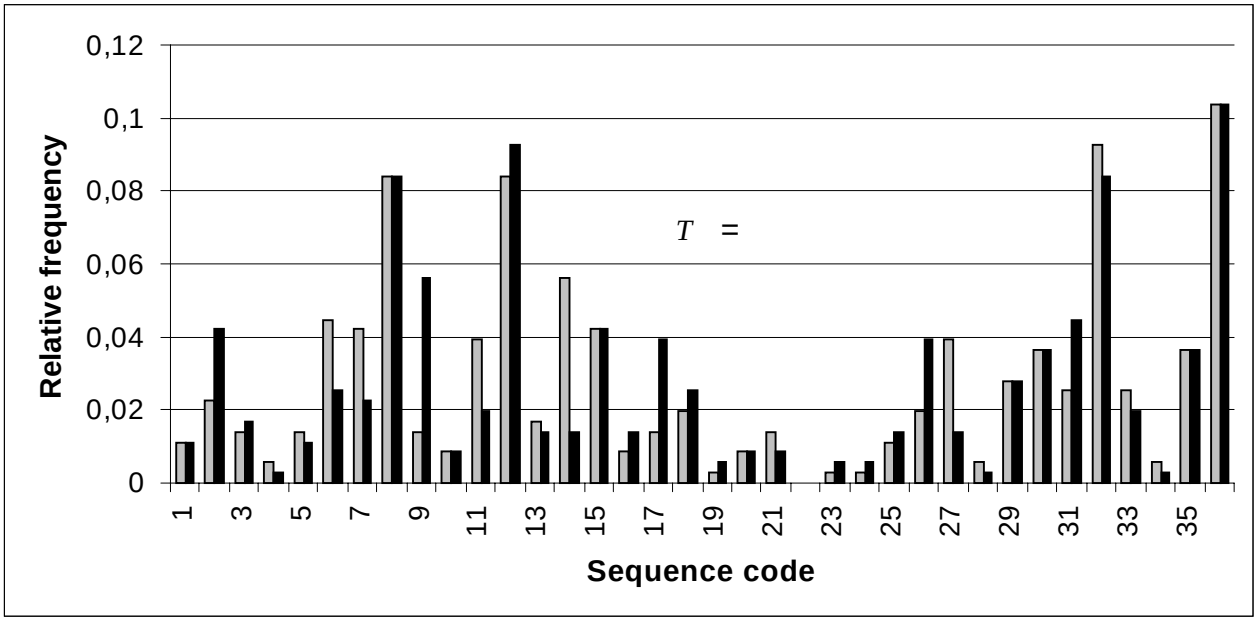


Figure 14: Forward and reverse-time symbol-sequence histograms for Italy. Six symbols and sequence length 2 where used. Forward-time histogram is plotted in black, reverse-time histogram is plotted in grey.

4.10 Conclusions

Through the above examples I have shown how STSA techniques can be used to characterize the nature of time patterns in data. It also appears to be a very effective method for detecting and quantifying time irreversibility. In particular STSA is useful to describe dynamics of economic models where the division in regimes plays an important role. The computational simplicity of the method suggests that it could be useful for the analysis of other types of economic data. It seems to me that future research in this area can include the use and development of prediction algorithms that with perform well when the data model is not known.

5 Evaluating the models – model fitting

We will consider the inverse problem of reconstructing a model of a dynamical system from measured time-series data. Such a model may be used for predictions, or control, or many other applications. Constructing models from time-series has a long tradition but substantial gaps in modeling capability remain. The method that we will describe tries to rescue substantial elements of our approach, where mainly we are interested in the (symbolic) dynamic across regimes. In our case the situation is the following: the system we wish to model is low dimensional and deterministic. However, no economic system can be isolated and are always subject to driving by various forms of noise. In addition, the observational data can be polluted by measurement noise. Decomposing a given signal as deterministic part plus noise is not a well-defined procedure without some strong constraints on the decomposing process. In most economic applications, these constraints are provided by the fact that one has some knowledge of the class of models to be considered both for the deterministic part and the noise. For the decomposition to be useful, the deterministic part should be as simple (and the noise be as small in amplitude) as possible. This can be made more precise by using a minimum description length criteria or some other objective measure to avoid overfitting.

We are interested in reconstruction of models using symbol sequence statistics. The observed symbol sequence statistics can be used as a target for measuring the goodness of fit of proposed models. In general, the reconstruction of dynamical systems from time-series data is done by varying the model dynamics such that some error function is minimized. This motivate the introduction of the error function E. A plot of this function vs. the model parameters constitutes the error landscape, and reconstruction of the dynamical system amounts to find the global minimum of this landscape. The error function E we have chosen is the Euclidean distance between the observed symbolic tree and the tree generated by the model. For our propose of finding a model that reproduce accurately the regime dynamics, this seems to be the better election of error function. Remember that we define the distance between two different tree branches of the same level m, A and B, as the Euclidean norm of their difference $d(A,B) = T_{AB}(m)$.

We denote the tree branch vector of level m measured from the data by $\boldsymbol{T}$. Choosing a set of parameters λ for our dynamical model $x_t = \varphi(x_{t-1};\lambda)$, we iterate the equation and calculate $\boldsymbol{T}(\lambda)$, the tree branch vector of level m for the model dynamics. The error function $E(\lambda)$ is a function on parameter space defined as

$$E(\lambda) = d(\boldsymbol{T}(\lambda),\boldsymbol{T}).$$

A plot of $E(\lambda)$ constitutes the error landscape. Reconstructing the dynamical model is equivalent to finding a global minimum in this landscape.[21]

5.1 Model development for the Italian economy

In this first experiment of fitting the model to the data we will only work with the Italian economy and the binary partition. To construct the error function E, we have to chose some fixed level of the tree and in our case the optimal value to compare is $m = 6$.[22] Our model is discrete in time and the dynamical variables which define the two-dimensional state space are the *growth rate of value added per person employed v* and the *growth rate of gross investment per person employed i*. The deterministic model can be represented by a first order system of two difference equations *(S)*: $z_t = T(z_{t-1})$, where $z = (v, i)$. System *(S)* can also be represented in the following way:

$$\begin{cases} v_t = T_1(v_{t-1}, i_{t-1}) \\ i_t = T_2(v_{t-1}, i_{t-1}) \end{cases},$$

where it is $T(v,i) = (T_1(v, i), T_2(v, i))$. Then the procedure to fit the model is the following. First we chose the functional form of the map ϕ that depends on a set of parameters λ. Then the initial condition of the time series data is iterated using the map T to produce a time series $s(\lambda)$ with the same length of the original that depends on λ. At this point, we select many values of λ for the artificial time series $s(\lambda)$ and transform these series to the corresponding binary symbolic sequences.[23] After computing the sixth level of the symbolic tree for each one of these symbolic sequences, the error function $E(\lambda)$ can be computed.[24] Finally, we compute the minimum λ_0 of the error function. This value give us the map T of the selected form whose dynamics better reproduce the observed regime dynamics. The results of the experiments are described in the following subsection.

5.2 Model fitting

We begin our experiments with a general piecewise-defined map, with a general form in each regime. In our case of two dynamical regimes, with Regime 0 = $\{(v,i) \in R^2 / v \geq 0, i \geq 0\}$ and Regime 1 = R^2 – Regime 0, this map can be written specifically by:

$$T : R^2 \to R^2$$

$$T(v,i) = \left(\begin{cases} T_1(v,i), \text{if } v \geq 0 \text{ and } i \geq 0 \\ S_1(v,i), \quad \text{otherwise} \end{cases}, \begin{cases} T_2(v,i), \text{if } v \geq 0 \text{ and } i \geq 0 \\ S_2(v,i), \quad \text{otherwise} \end{cases} \right)$$

---

[21] See X.Z. Tang, E.R. Tracy, A.D. Boozer and R. Brown (1995) for an application of these techniques in examining the error landscape for the logistic, Ikeda and Hénon maps. They found that the approach is highly robust even in the presence of observational and dynamical noise. They also showed that, even in the case where the noise is comparable to the signal, it is possible by including the noise characteristics as part of the model to produce robust considerations.

[22] See section 4.4 of this paper.

[23] In general, the error function $E(\lambda)$ cannot be computed for a generic value of λ. It is for this reason that we compute it for many different values and employ conventional routines to find the global minimum.

[24] The search range for the parameter vector λ is not specified and we used an ad oc selection. Only those results that passed an error condition $E(\lambda) \leq \varepsilon$ are retained, where $\varepsilon > 0$ is a given value. In our experiments, we used $\varepsilon = 0{,}01$.

The first step is experimenting with simple maps T having (0,0) as fixed point. Three different target maps are used, one with piecewise linear component functions and the others with quadratic and linear components. In each case, the set of parameters λ is represented by the coefficients of the map T.

1) Piecewise linear:

$$T : R^2 \to R^2$$

$$T(v,i) = \left(\begin{cases} a_1 v + a_2 i, & \text{if } v \geq 0 \text{ and } i \geq 0 \\ a_3 v + a_4 i, & \text{otherwise} \end{cases}, \begin{cases} b_1 v + b_2 i, & \text{if } v \geq 0 \text{ and } i \geq 0 \\ b_3 v + b_4 i, & \text{otherwise} \end{cases} \right)$$

Results:

$\{ a_1 = .1967199028 , a_2 = .08206942038 , a_3 = .2897322649, a_4 = -.2779930329\}$

$\{ b_1 = .3508386324 , b_2 = .1263851405, \ b_3 = .4767596907, b_4 = -.08748268761\}$

$E(\lambda) = 0{,}00410$

2) Quadratic in *v*:

$$T : R^2 \to R^2$$

$$T(v,i) = \left(\begin{cases} a_1 v^2 + a_2 i, & \text{if } v \geq 0 \text{ and } i \geq 0 \\ a_3 v^2 + a_4 i, & \text{otherwise} \end{cases}, \begin{cases} b_1 v^2 + b_2 i, & \text{if } v \geq 0 \text{ and } i \geq 0 \\ b_3 v^2 + b_4 i, & \text{otherwise} \end{cases} \right)$$

Results:

$\{ a_1 = .005956983987, a_2 = .1275047297, a_3 = .004543611267, a_4 = -.2679357916\}$

$\{ b_1 = .01772310954 , b_2 = .1807632846, \ b_3 = -.001204062036, b_4 = -.1017623345\}$

$E(\lambda) = 0{,}00362$

3) Quadratic in *i*:

$$T : R^2 \to R^2$$

$$T(v,i) = \left(\begin{cases} a_1 v + a_2 i^2, & \text{if } v \geq 0 \text{ and } i \geq 0 \\ a_3 v + a_4 i^2, & \text{otherwise} \end{cases}, \begin{cases} b_1 v + b_2 i^2, & \text{if } v \geq 0 \text{ and } i \geq 0 \\ b_3 v + b_4 i^2, & \text{otherwise} \end{cases} \right)$$

Results:

{ a_1 = .2536302448, a_2 = .001811070891, a_3 = .3753043824, a_4 = .009010593092}

{ b_1 = .3820517995, b_2 = .004832264929, b_3 = .5061705486, b_4 = .003172125965}

$E(\lambda)$ = 0,00314

4) General piecewise linear in each regime

We finish our experiments with a general piecewise linear map of the plane, defined by

$$T : R^2 \to R^2$$

$$T(v,i) = \left(\begin{cases} a_1 v + a_2 i + a_3 \text{ , if } v \geq 0 \text{ and } i \geq 0 \\ a_4 v + a_5 i + a_6 \text{ , otherwise} \end{cases} , \begin{cases} b_1 v + b_2 i + b_3 \text{ , if } v \geq 0 \text{ and } i \geq 0 \\ b_4 v + b_5 i + b_6 \text{ , otherwise} \end{cases} \right)$$

This map is linear in each regime and in this case we do not impose that (0, 0) is a fixed point of *T*.

Results:

{ a_1 = .06641527724, a_2 = .02221151630, a_3 = 2.077066861, a_4 = .2620545325, a_5 = -.2219829557, a_6 = 1.864989192}

{ b_1 = .4566736879, b_2 = .1750024830, b_3 = -1.687019825, b_4 = .4662332857, b_5 = -0661809151, b_6 = .7092933519}

$E(\lambda)$ = 0,00177

Note that in this case we have obtained the best representation of the observed regime dynamics of the four experiments. This fact is reflected by the values of the error function: in this case the simulated symbolic tree is closer to the observed one.

5.3 Conclusions

In this section I have addressed the problem of (black box) reconstruction of dynamical models from data sets using the symbol sequence statistics. In particular, I have showed how to compute the parameter set of the model using the symbol sequence statistics as the target. The method seems to be the correct direction if the objective is reproducing the dynamics across regimes of the economic system. I want to emphasize again the experimental and partial character of the exercises reproduced above. I am far from having obtained a decent model and there is still a lot to do for satisfactory results. The reason why I experimented with the two regimes model and with simple forms of map *T* is that the computer CPU time required for convergence of the calculations becomes prohibitive as the number of fitting parameters increases. It seems to me that to go ahead requires the use and extension of these results, the research of optimal symbolic representations of data and the development of new computational algorithms.

References

J.S. Armfield, C.S. Daw, P. Durbetaki, C.E.A. Finney and J.B.Green (1998), "Time Irreversibility of Cycle-by-Cycle Engine Combustion Variations", Proceedings of the *1998 Technical Meeting of the Central States Section of the Combustion Institute*.

J.S. Armfield, C.S. Daw, J.A. Dralimeier, P. Durbetaki, C.E.A. Finney, J.B.Green, M.B. Kennel and R.M. Wagner (1998), "Time Irreversibility and Comparison of Cyclic-Variability Models", To appear at the *1999 SAE International Congress & Exposition*.

Böhm, B. and L. F. Punzo (1992), "Detecting Structural Change: A Scheme for the Comparison of Austria and Italy in the Seventies and Eighties", in O. Clauser, P. Kalambach, G. Pergoretti and M. L. Segana (eds.), *Technological Innovation, Competitivness and Economic Growth*, Dunker & Humbolt, Berlin.

Böhm B. (1996), "Dynamic Econometric Specification and the Analysis of Structural Change", *Dynamis*, Quaderno 2/96, IDSE, Milano.

Böhm, B. and L. F. Punzo (1994): "Dynamics of Industrial Sectors and Structural Change in the Austrian and Italian Economies, 1970-1989", in Böhm B. and L. F. Punzo (eds.), *Economic Performance. A Look at Austria and Italy*, Physica Verlag, Heidelberg.

Böhm, B. and L. F. Punzo (1995), "Structural Change in the Context of Uneven Regional Development. The Path of Italian Dualistic Economy Revisited with a New Dynamical Approach", paper presented at the XIth International Conference on Input Output Techniques, New Delhi, 1995.

F.T. Connolly, C.S. Daw, C.E.A. Finney and M.B. Kennel (1998), "Observing and Modelling Non-linear Dynamics in an Internal Combustion Engine", *Physical Review E* **57**:3, 2811-2819.

J.P. Cruchfield and N.H. Packard (1983), "Symbolic Dynamics of Noisy Chaos", *Physica D* **7**, 201-223.

J.P. Cruchfield and K. Young (1989), "Infering Statistical Complexity", *Physical Review Letters* **63**:2, 105-108.

C.S. Daw, C.E.A. Finney and J.B.Green (1998), "Symbolic Time-Series Analysis of Engine Combustion measurements", SAE Paper No. 980624.

C.S. Daw, C.E.A. Finney, K. Nguyen and J.S. Halow (1998a), "Symbolic Statistics: A New Tool for Understanding Multiphase Flow Phenomena", presented at the *1998 ASME International Congress & Exposition*.

C.S. Daw, C.E.A. Finney, K. Nguyen and J.S. Halow (1998b), "Symbolic-Sequence Statistics for Monitoring Fluidization", *1998 International Mechanical Engineering Congress & Exposition*.

C.S. Daw, K.D. Edwards, C.E.A. Finney and K. Nguyen (1998), "Use of Symbolic Statistics to Characterize Combustion in a Pulse Combustor Operating Near the Fuel-Lean Limit", Proceedings of the *1998 Technical Meeting of the Central States Section of the Combustion Institute*.

C. Diks, J.C. van Houwelingen, F. Takens, J. DeGoede (1995), “Reversibility as a Criterion for Discriminating Time Series”, *Physics Letters A* **201**, 221-228.

Hicks, J. (1973), *Capital and Time*, Oxford University Press, Oxford.

J. Kurths, U. Schwarz, A. Witt, R. Th. Krampe and M. Abel (1996), "Measures of complexity in signal analysis", in Chaotic, Fractal, and Nonlinear Signal Processing, AIP Conference Proceedings 375 (3rd Technical Conference on Nonlinear Dynamics and Full Spectrum Processing (Mystic, Connecticut USA;
1995 July 10-14) Woodbury, New York 1996, 33-54.

M. Lehrman, A.B. Rechester and R.B. White (1997), “Symbolic Analysis of Chaotic Signals and Turbulent Fluctuations”, *Physical Review Letters* **78**:1, 54-57.

Punzo, L. F. (1995), “Some Complex Dynamics for a Multisectoral Economy”, *Revue Economique*.

C. Shannon (1963), *A Mathematical Theory of Communication*, Bell System Technical Journal **27**, 379-473 and 623-656, 1948, reprinted in *The Mathematical Theory of Communication* by C. Shannon and W. Weaver, University of Illinois Press.

X.Z. Tang and E.R. Tracy (1997), “Data Compression and Information Retrieval via Symbolization”, *Chaos* **8**:3, 688-696.

X.Z. Tang, E.R. Tracy and R. Brown (1997), “Symbol Statistics and Spatio-Temporal Systems”, *Physica D* **102,** 253-261.

X.Z. Tang, E.R. Tracy, A.D. Boozer and R. Brown (1995), “Symbol Sequence Statistics in Noisy Chaotic Signal Reconstruction”, *Physical Review E* **51**:5, 3871-3889.

X.Z. Tang, E.R. Tracy, A.D. Boozer, A. deBrown and R. Brown (1994), “Reconstruction of Chaotic Signal Using Symbolic Data ”, *Physical Letters A* **190**, 393-398.

H. Voss and J. Kurths (1998), “Test for Nonlinear Dynamical Behaviour in Symbolic Sequences”, *Physical Review E* **58**:1, 1155-1158.

G. Weiss (1975), “Time-Reversibility of Linear Stochastic Processes”, *Journal of Applied Probability* **12**, 831-836, 1975.

A. Witt, J. Kurths, A.O. Benz and U. Schwarz (1993), “Analysis of Solar Spike Events by Means of Symbolic Dynamics Methods”, *Astronomy and Astrophysics* **227**, 215-224.